UNIVERSITY OF CALIFORNIA
Los Angeles

Topics in Geometric and Topological Data Analysis

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Mathematics

by

Abigail Hickok

2023

ABSTRACT OF THE DISSERTATION

Topics in Geometric and Topological Data Analysis

by

Abigail Hickok
Doctor of Philosophy in Mathematics
University of California, Los Angeles, 2023
Professor Mason Alexander Porter, Chair

The fields of topological data analysis (TDA) and geometric data analysis (GDA) use algebraic topology and differential geometry to capture topological and geometric structural properties of data that are not captured by other methods in data science and machine learning. The primary tool of TDA—and one of the focuses of this dissertation—is persistent homology, which measures the connected components, holes, and higher-dimensional voids of a data set and tracks how those voids emerge and disappear at different scales. The objective of GDA is to extract new insights by considering geometric invariants of a manifold, such as curvature, rather than topological invariants. Previous studies have demonstrated the power of geometry and topology for analyzing data in complex systems, neuroscience, biology, and many other fields.

In my thesis, I study both the theory and applications of topological and geometric data analysis. In the first part of the dissertation, I establish and analyze a new construction, called a "persistence diagram (PD) bundle," for doing multiparameter TDA, and I develop an algorithm to compute a certain class of PD bundles. PD bundles generalize several important constructions in TDA: vineyards, the persistent homology transform, and fibered

barcodes. In the second part of the dissertation, I apply TDA to several geospatial and geospatiotemporal data sets. In the last part of the dissertation, I introduce a new method for curvature estimation in point-cloud data.

The dissertation of Abigail Hickok is approved.

Christopher R. Anderson

Michael Anthony Hill

Deanna M. Hunter

Mason Alexander Porter, Committee Chair

University of California, Los Angeles

2023

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Mason Porter. I started graduate school with a background only in pure mathematics, and it was mainly Mason who taught me how to be an *applied* mathematician. More importantly, Mason has been extremely generous with his time and has supported my career in innumerable ways that go beyond what is expected of an advisor. He has been a wonderful mentor throughout graduate school, and I would not be where I am without him.

In addition to Mason, I am especially grateful for the mentorship of Andrew Blumberg. During my visit to Columbia last year, I learned so much from him about geometric data analysis and applications in biology. His support and feedback on my postdoc applications was also immensely helpful. I look forward to many more collaborations during my next few years as a postdoc.

I'd like to express my great appreciation for my co-authors and the many other mathematicians whose insights and comments contributed directly to the contents of this thesis. Chapter 3 benefited immensely from conversations with Andrew Blumberg, Ryan Grady, and Karthik Viswanathan. I thank Michael Lesnick and Nina Otter for helpful discussions about the contents of Chapter 4. Chapter 5 is adapted from a paper [HNP22] that I co-authored with Deanna Needell and Mason Porter, and I thank Henry Adams, Heather Zinn Brooks, Michelle Feng, Lara Kassab, and Nina Otter for helpful discussions. Chapter 6 is adapted from a paper [HJJ22] that I co-authored with Benjamin Jarman, Michael Johnson, Jiajie Luo, and Mason Porter; I thank Chris Anderson and Renata Turkes for helpful comments and discussions. Chapter 7 is adapted from a paper in preparation that is co-authored with Andrew Blumberg. Conversations with Mike Hill about the theoretical sections of this thesis were also very helpful. Lastly, I thank Mason Porter for providing comments and editing advice on every chapter of this thesis.

I would like to thank my undergraduate professors, Peter Ozsváth and Zoltán Szabó, for

inspiring my initial interest in topology. I am grateful that they believed in me and helped push me towards success in graduate school.

I thank my committee members Michael Hill, Deanna Needell, and Chris Anderson for their insightful comments and questions on the work in this thesis. I also learned lots of mathematics from all of them during the courses that I took with them early in graduate school. In addition to their mathematical insights, I am especially thankful to Mike Hill and Deanna Needell for writing recommendation letters for me to support the next step of my academic career.

I would also like to thank the other members and alumni of my research group, who often provided mentorship and helpful feedback on projects. I give a special thanks to Michelle Feng, Gill Grindstaff, Michael Johnson, Yacoub Kureh, Grace Li, Jerry Luo, Nina Otter, and Sarah Tymochko.

I was fortunate to receive funding throughout my PhD from the UCLA Eugene V. Cota-Robles Fellowship, the UCLA Graduate Dean's Scholar Fellowship, the UCLA Dissertation Year Fellowship, and the National Science Foundation (grant number 1922952) through the Algorithms for Threat Detection (ATD) program.

I am deeply grateful for my friends and family and their unwavering support throughout my academic journey. My parents' belief in me has been instrumental in my success, and the encouragement of my siblings, Nano and Eleanor, has kept me motivated. I also thank my extended family for inspiring my curiosity, and I especially thank my grandma, who is not only an inspiration for getting her own PhD in mid-life, but was also a wonderful "roommate" for several months of graduate school. This accomplishment would not have been possible without all of them.

Lastly, I would like to express my heartfelt gratitude to my partner, Josh, for his limitless love and support, for keeping me sane this year during application season and while finishing this dissertation, and especially for agreeing to move across the country to continue this

academic journey with me. I would not have been able to do it without him.

VITA

2014–2018 A.B. Mathematics, Princeton University.

PUBLICATIONS AND PREPRINTS

Abigail Hickok. "Computing Persistence Diagram Bundles." *arXiv:2210.06424*, 2022.

Abigail Hickok. "Persistence Diagram Bundles: A Multidimensional Generalization of Vineyards." *arXiv:2210.05124*, 2022.

Abigail Hickok, Benjamin Jarman, Michael Johnson, Jiajie Luo, Mason A. Porter. "Persistent Homology for Resource Coverage: A Case Study of Access to Polling Sites." *arXiv:2206.04834*, 2022.

Abigail Hickok. "A Family of Density-Scaled Filtered Complexes." *arXiv:2112.03334*, 2022.

Abigail Hickok, Deanna Needell, Mason A. Porter. "Analysis of Spatial and Spatiotemporal Anomalies Using Persistent Homology: Case Studies with COVID-19 Data." *SIAM Journal on Mathematics of Data Science*, 4(3):1116–1144, 2022.

Michelle Feng, Abigail Hickok, Mason A. Porter. "Topological Data Analysis of Spatial Systems." In Federico Battiston and Giovanni Petri (eds.) *Higher-Order Systems*, ch. 17, pp. 389–399. Springer, Cham, Switzerland, 2022.

Abigail Hickok, Yacoub H. Kureh, Heather Zinn Brooks, Michelle Feng, Mason A. Porter. A Bounded-Confidence Model of Opinion Dynamics on Hypergraphs. *SIAM Journal on Applied Dynamical Systems.* 21(1):1—32, 2022.

Christian R. Scullard, Abigail Hickok, Justyna O. Sotiris, Bilyana M. Tzolova, Loek Van Heyningen, Frank R. Graziani. "Adaptive Spectral Solution Method for the Landau and Lenard-Balescu Equations." *Journal of Computational Physics* 402, 109110, 2020.

CHAPTER 1

Introduction

It is common for high-dimensional data sets and systems to have lower-dimensional manifold structure, making them more amenable to analysis. Recently, the fields of topological data analysis (TDA) and geometric data analysis (GDA) have become popular for studying the topology, geometry, and manifold structure of data. In TDA, one uses algebraic topology to analyze the "shape" and global structure of (typically) high-dimensional data [Car20]. These structural properties are often ignored by traditional data-science methods, such as dimension-reduction algorithms. GDA concerns the inference of geometric invariants of a manifold, such as curvature, from discrete samples. In addition to its intrinsic mathematical value, GDA also has the potential to improve manifold-learning algorithms, in which one seeks to infer lower-dimensional manifold structure of high-dimensional data. This thesis explores new methods, theory, and applications of TDA and GDA.[1]

In my research on TDA, I focus on *persistent homology* (PH), which is an approach from algebraic topology that one can use to algorithmically find holes of different dimensions in a data set and quantify the "persistence" of those holes [OPT17]. The computation of PH has yielded insights into a wide variety of areas, including neuroscience [SPG19, GGB16], materials science [BHO18], chemistry [MTC10], biology [RB19], dynamical systems [MZR16, YB20], and collective behavior [TZH15b]. Analyzing PH allows one to quantify holes in data in a meaningful way and has made it possible to apply topological ideas to a wide variety of empirical data sets [OPT17].

[1]The remainder of this introduction, as well as the rest of the thesis, is adapted from [HNP22, Hic22b, Hic22c, Hic22a, HJJ22, FHP22, HB23].

In algebraic topology, homology is a quantitative way to characterize a topological space. For example, suppose that M is a compact, orientable manifold. The rank of the 0-dimensional homology $H_0(M)$ is the number of connected components, the rank of $H_k(M)$ is the number of k-dimensional holes in M for $k \geq 1$, and the dimension of M is equal to the largest integer n such that $H_n(M)$ is nontrivial[2]. For instance, if M is the torus $S^1 \times S^1$, then there is one connected component, there are two 1-dimensional holes, and there is one 2-dimensional hole. Although homology does not uniquely identify a topological space, it provides useful information about a space's global structure, and homology can be used to distinguish a topological space from other topological spaces that have different homology groups.

In TDA, data often takes the form of a set X of points in $\mathbb{R}^n$, called a *point cloud*, and we hope to analyze the topology of a (hopefully lower-dimensional) subspace that the points lie on. The first step of PH is to construct a *filtered simplicial complex*, which is a combinatorial description of a topological space (see Definition 2.2.1). One of the classical approaches to building a filtered complex is the *Čech complex* $\check{\mathrm{C}}(X)$ (see Definition 2.2.2). At each point $x \in X$, one places a ball of radius $r > 0$, where r is the *filtration parameter*, and then adds a k-simplex with vertices $X' \subseteq X$ to $\check{\mathrm{C}}(X)_r$ if the intersection $\bigcap_{x \in X'} B(x, r)$ is nonempty. For an example, see Figure 1.1. The PH of $\check{\mathrm{C}}(X)$ records how the homology of $\check{\mathrm{C}}(X)_r$ changes as r increases. As r grows, new homology classes (which represent k-dimensional holes) are "born" and old homology classes "die". One way to summarize this information is with a *persistence diagram* (PD), which is a multiset of points in the extended plane $\overline{\mathbb{R}}^2$ (see Section 2.3). If there is a homology class that is born at filtration-parameter value b and dies at filtration-parameter value d, the PD contains the point (b, d). By examining a PD, we can see how the topology of a data set differs at different scales.

In my research on GDA, I focus on estimating curvature from point clouds sampled from Riemannian manifolds. A *Riemannian manifold* is a smooth manifold M equipped

[2]The homology group $H_n(M)$ is "nontrivial" when there are n-dimensional holes.

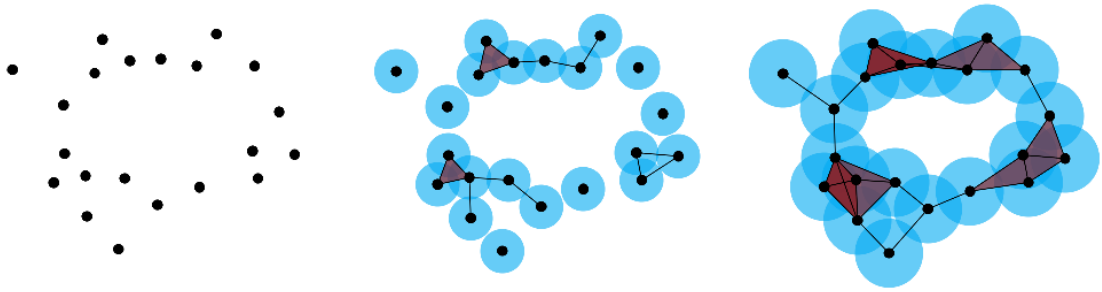


Figure 1.1: The Čech filtration for a point cloud that is sampled from an annulus. Initially, all the vertices are disconnected, so there are many 0D homology classes. As the filtration parameter r increases, larger connected components form and 0D homology classes die. At an even larger filtration-parameter value, a 1D hole appears; that is, a 1D homology class is born. (This figure appeared originally in [HJJ22].)

with compatible choices of inner product for each tangent space T_pM. The presence of this structure equips M with a metric (which is induced by the fact that the inner product allows one to define the length of a tangent vector) and moreover lets us makes sense of various geometric notions on M—in particular, the notion of *curvature*. Addionally, many methods have been developed to estimate other quantities in differential geometry. For example, estimating *geodesic distance* (which is the distance between two points on a Riemannian manifold) is an especially important task, and we rely on methods to estimate it in our work on curvature estimation.

1.1 Contributions of the thesis

In the first part of this thesis, I introduce the concept of a persistence diagram (PD) bundle, which is the space of PDs for a fibered filtration function (a set $\{f_p : \mathcal{K}^p \to \mathbb{R}\}_{p\in B}$ of filtrations that is parameterized by a topological space B). Special cases include vineyards [CEM06], which can be visualized as continuously-varying stacks of PDs (see Figure 1.2 and Section 2.5), the persistent homology transform [TMB14], and fibered barcodes for multiparameter

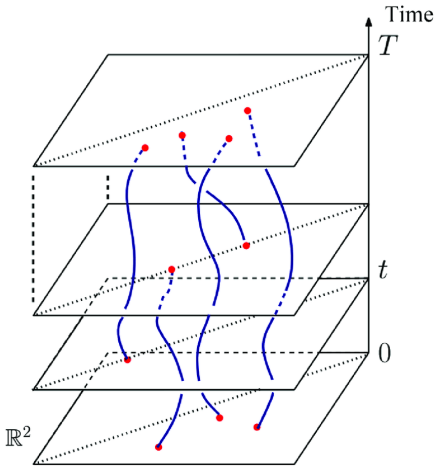


Figure 1.2: An illustration of a vineyard, which consists of a persistence diagram for each time t. (This figure is a slightly modified version of a figure that appeared originally in [LWA17]; the figure is available under a Creative Commons license.)

persistence modules [And13]. I prove that if B is a smooth compact manifold, then for a generic fibered filtration function, B is stratified such that within each stratum $Y \subseteq B$, there is a single PD "template" (a list of "birth" and "death" simplices) that can be used to obtain the PD for the filtration f_p for any $p \in Y$. If B is compact, then there are finitely many strata, so the PD bundle for a generic fibered filtration on B is determined by the persistent homology at finitely many points in B. I also show that not every local section (a continuous map from an open set in B to the total space E of PDs) can be extended to a global section (a continuous map $B \to E$). Consequently, the points in the PDs do not typically trace out separate manifolds as $p \in B$ varies; this is unlike a vineyard, in which the points in the PDs trace out curves ("vines"). When there is a stratification as described above, I construct a cellular sheaf that stores sufficient data to construct sections and determine whether a given local section can be extended to a global section.

I also develop an algorithm to compute "piecewise-linear" PD bundles, a wide class that includes many of the PD bundles that one encounters in practice. I give full details for the case in which B is a triangulated surface, and I outline generalizations to higher dimensions

and other cases. The algorithm relies on the stratification that I described above.

In the second part of the thesis, I apply TDA in studies of geospatial data. In Chapter 5, we analyze local extrema of real-valued spatial and spatiotemporal data. Our TDA approach captures both local information (specifically, the locations and values of the local extrema) and global information about the relationships between the extrema, such as the extent to which extrema are "spatially separated". By contrast, prior methods to analyze local extrema yielded only local information. We conduct two case studies: first, we examine COVID-19 vaccination rates in New York City by zip code at a single point in time; second, we study a year-long data set of COVID-19 case rates in neighborhoods of Los Angeles. We use vineyards to track how the local and global structure of the extrema changed with time.

In Chapter 6, we use TDA as a framework to measure and evaluate the equity of the geographic distribution of a resource (e.g., polling places or vaccination sites). Our aim is to identify underserved geographic regions ("holes in coverage"). We use PH because it allows us to study holes in coverage at all scales without needing to choose an arbitrary cutoff distance. Instead of using geographic distance, we use travel time, which better captures the "cost" of accessing a resource site. We additionally incorporate wait times at the resource sites using a weighted Vietoris–Rips filtration. As a case study, we analyze access to polling places and use PH to identify holes in the distribution of polling places.

In the last part of the thesis, we introduce an intrinsic estimator for the scalar curvature of a data set that is presented as a finite metric space. Such data sets include point clouds (from which geodesic distances can be estimated) and networks that are equipped with the shortest-path metric. Our estimator depends only on the metric structure of the data and not on an embedding in $\mathbb{R}^n$. We show that the estimator is consistent in the sense that for points sampled from a probability measure on a compact Riemannian manifold, the estimator converges to the scalar curvature as the number of points increases. To justify its use in applications, we show that the estimator is stable with respect to perturbations of the metric structure. We validate our estimator experimentally on synthetic data that is

sampled from manifolds with known curvature.

1.2 Organization

The rest of the thesis is organized as follows. I review background material on TDA in Chapter 2. In Chapters 3 and 4, I present my research on PD bundles; these chapters are adapted from [Hic22c] and [Hic22a], respectively. My research on geospatial applications of TDA is in Chapters 5 and 6; these chapters are adapted from [HNP22], [FHP22], and [HJJ22]. In Chapter 7, I discuss curvature estimation; this chapter is adapted from [HB23]. Finally, in Appendix A, I briefly mention my research on opinion dynamics [HKF22], which I do not include in the dissertation for narrative reasons.

CHAPTER 2

Background

In this chapter, I review the fundamentals of TDA and GDA.[1]

2.1 Topological data analysis

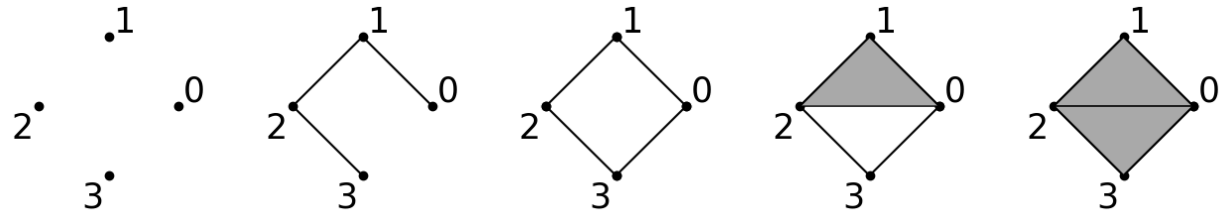
See [EH10, Ghr14, OPT17] for comprehensive introductions to TDA.

2.2 Filtered complexes

In TDA, one usually works with simplicial complexes, a certain combinatorial description of a topological space, rather than general topological spaces. Simplicial complexes are defined as follows. A *k-simplex* σ is the k-dimensional convex hull of $k+1$ vertices. The convex hull of a nonempty subset of these vertices is a *face* of the simplex. If τ is a face of σ and $\tau \neq \sigma$, then τ is a *proper face* of σ. A *simplicial complex* $\mathcal{K}$ is a set of simplices that satisfies two requirements: (1) if $\sigma \in \mathcal{K}$ is a simplex, then every face of σ is also a simplex in $\mathcal{K}$; (2) if σ and τ are simplices in $\mathcal{K}$, then $\sigma \cap \tau$ is a face of both σ and τ. Constraining ourselves to simplicial complexes allows us to reduce the computation of homology (and persistent homology, defined below in Section 2.3) to a linear-algebra calculation.

Consider a simplicial complex $\mathcal{K}$. A *filtration function* $f : \mathcal{K} \to \mathbb{R}$ is a real-valued function on $\mathcal{K}$ that is *monotonic*. That is, $f(\tau) \leq f(\sigma)$ if τ is a face of σ. Monotonicity guarantees

[1] This chapter is adapted from [Hic22c, Hic22a, Hic22b].



(a) $\mathcal{K}_0$ (b) $\mathcal{K}_1$ (c) $\mathcal{K}_2$ (d) $\mathcal{K}_3$ (e) $\mathcal{K}_4$

Figure 2.1: An example of a filtered complex $\{\mathcal{K}_r\}_{r\in\mathbb{R}}$. We show the simplicial complex $\mathcal{K}_r$ at $r = 0, 1, 2, 3, 4$. (This figure appeared originally in [HNP22].)

that the r-sublevel sets $\mathcal{K}_r := \{\sigma \in \mathcal{K} \mid f(\sigma) \leq r\}$ are simplicial complexes. The set $\{\mathcal{K}_r\}_{r\in\mathbb{R}}$ is a *filtered complex*.

Definition 2.2.1. A *filtered complex* is a set $\{\mathcal{K}_r\}_{r\in\mathbb{R}}$ of simplicial complexes such that $\mathcal{K}_r \subseteq \mathcal{K}_s$ whenever $r \leq s$.

The *filtration value* of a simplex $\sigma \in \mathcal{K}$ is $f(\sigma)$. The parameter r is the *filtration parameter*. We will sometimes refer to the pair $(\mathcal{K}, f)$ as the filtered complex itself. In Figure 2.1, we show an example of a filtered complex.

We review the standard methods for building a filtered complex from point cloud data. For the remainder of this dissertation, let $(\mathbb{X}, d)$ denote a metric space, let $X = \{x_1, \ldots, x_N\}$ denote a point cloud in $\mathbb{X}$, and let $d_X = d|_X$. Often $(\mathbb{X}, d)$ is a Euclidean metric space, but there are exceptions, such as in Chapter 6. For any index set $J \subseteq \{1, \ldots, N\}$, let x_J denote the simplex with vertices x_j for all $j \in J$.

Definition 2.2.2. The *Čech filtered complex* $\check{\mathrm{C}}(\mathbb{X}, d, X)$ is the filtered complex such that the simplicial complex $\check{\mathrm{C}}(\mathbb{X}, d, X)_r$ consists of the simplices

$$\left\{x_J \mid \bigcap_{j\in J} B(x_j, r) \neq \emptyset \text{ and } J \subseteq \{1, \ldots, N\}\right\},$$

where $B(x, r) := \{y \in \mathbb{X} \mid d(x, y) \leq r\}$. (Equivalently, $\check{\mathrm{C}}(\mathbb{X}, d, X)_r$ is the nerve of $\{B(x, r)\}_{x\in X}$.)

The Nerve Theorem provides theoretical guarantees for the Čech complex.

Theorem 2.2.3 (Nerve Theorem [Bor48])**.** *If $\bigcap_{j\in J} B(x_j, r)$ is either contractible or empty for all $J \subseteq \{1, \ldots, N\}$, then $\check{C}(\mathbb{X}, d, X)_r$ is homotopy-equivalent to $\bigcup_{i=1}^{N} B(x_i, r)$.*

Homotopy-equivalence is important because homotopy-equivalent spaces have the same homology.

The Nerve Theorem must be applied with care. In Euclidean space, all balls are convex (hence their intersections are contractible), so the simplicial complex $\check{\mathrm{C}}(\mathbb{X}, d, X)_r$ is homotopy-equivalent to $\bigcup_{x\in X} B(x, r)$ for all r. In other metric spaces, however, balls are not always convex. In a Riemannian manifold, $\bigcap_{j\in J} B(x_j, r)$ is contractible when r is smaller than the convexity radius, but even individual balls $B(x, r)$ may not be contractible when r is too large. For example, on a sphere S^n, a closed ball of radius π is the entire sphere, which is not contractible. When r is large, the Nerve Theorem may not be applicable.

Computing the Čech filtered complex is computationally intensive [OPT17]. In practice, researchers often compute the Vietoris–Rips filtered complex instead, which requires only pairwise distances between the points in X.

Definition 2.2.4. The *Vietoris–Rips filtered complex* $\mathrm{VR}(d_X)$ is the filtered complex such that the simplicial complex $\mathrm{VR}(d_X)_r$ consists of the simplices

$$\Big\{ x_J \mid d_X(x_i, x_j) \leq 2r \text{ for all } i, j \in J \text{ and } J \subseteq \{1, \ldots, N\} \Big\}.$$

At each r, the simplicial complexes $\mathrm{VR}(d_X)_r$ and $\check{\mathrm{C}}(\mathbb{X}, d, X)_r$ share the same 1-skeleton (the edges of the simplicial complexes). When the metric space $(\mathbb{X}, d)$ is Euclidean, the Vietoris–Rips filtered complex and the Čech filtered complex are related by the Vietoris–Rips lemma [EH10], which says that

$$\check{\mathrm{C}}(\mathbb{X}, d, X)_r \subseteq \mathrm{VR}(d_X)_r \subseteq \check{\mathrm{C}}(\mathbb{X}, d, X)_{\sqrt{2}r}$$

for all r.

Weighted versions of the Čech and Vietoris–Rips filtered complexes were introduced in [BLM19]. We recall the definitions here.

Definition 2.2.5. Let $\mathcal{C}^1_+([0,\infty))$ denote the set of differentiable bijective functions $\phi : [0,\infty) \to [0,\infty)$ with positive first derivative. A *radius function* on X is a function $r : X \to \mathcal{C}^1_+([0,\infty))$. For all points x in X, we denote $r(x) \in \mathcal{C}^1_+([0,\infty))$ by r_x.

The protypical example is the radius function of the form

$$r_x(t) := \begin{cases} -\infty\,, & t < w_x \\ t - w_x\,, & \text{otherwise,} \end{cases} \tag{2.1}$$

where $\{w_x\}_{x\in X}$ (the set of *weights*) is a set of real numbers. In this case, the ball $B(x, r_x(t))$ has no points until $t = w_x$; at that time, the radius starts growing linearly with t. Another example is the radius function $r_x(t) := ts_x$, where $\{s_x\}_{x\in X}$ is a set of positive real numbers. This is the case in which a ball centered at $x \in X$ grows linearly with t at rate s_x.

Definition 2.2.6. Let $r : X \to \mathcal{C}^1_+([0,\infty))$ be a radius function. The *weighted Čech filtered complex* $\check{\mathrm{C}}^{\mathrm{weighted}}(\mathbb{X}, d, X, r)$ is the filtered complex such that the simplicial complex $\check{\mathrm{C}}^{\mathrm{weighted}}(\mathbb{X}, d, X, r)_t$ consists of the simplices

$$\left\{ x_J \mid \bigcap_{j\in J} B(x_j, r_{x_j}(t)) \neq \emptyset \text{ and } J \subseteq \{1,\dots,N\} \right\},$$

where t denotes the filtration parameter. Equivalently, $\check{\mathrm{C}}^{\mathrm{weighted}}(\mathbb{X}, d, X, r)_t$ is the nerve of $\{B(x, r_x(t))\}_{x\in X}$. When the radius function is of the form in Equation 2.1, we use the notation $\check{\mathrm{C}}^{\mathrm{weighted}}(\mathbb{X}, d, X, \{w_x\})$.

Definition 2.2.7. Let $r : X \to \mathcal{C}^1_+([0,\infty))$ be a radius function. The *weighted Vietoris–Rips filtered complex* $\mathrm{VR}^{\mathrm{weighted}}(d_X, r)$ is the filtered complex such that the simplicial complex $\mathrm{VR}^{\mathrm{weighted}}(d_X, r)_t$ consists of the simplices

$$\left\{ x_J \mid d_X(x_i, x_j) \leq r_{x_i}(t) + r_{x_j}(t) \text{ for all } i, j \in J \text{ and all } J \subseteq \{1,\dots,N\} \right\},$$

where t denotes the filtration parameter. When the radius function is of the form in Equation 2.1, we use the notation $\mathrm{VR}^{\mathrm{weighted}}(d_X, \{w_x\})$.

In addition to the Čech and Vietoris–Rips filtered complexes (and their weighted versions), there are many other methods for constructing a filtered complex from a point cloud [OPT17]. For example, in the k-nearest neighbors filtered complex, each point is connected to its k-nearest neighbors at the kth filtration step.

Definition 2.2.8. The *k-nearest neighbors* filtered complex $\mathrm{KNN}(d_X)$ is the filtered complex such that $\mathrm{KNN}(d_X)_k$ is the simplicial complex that consists of the simplices

$$\Big\{ x_J \mid d_X(x_i, x_j) \leq d_X(x_i, x_{N_i^k}) \text{ or } d_X(x_i, x_j) \leq d_X(x_i, x_{N_j^k}) \\ \text{for all } i, j \in J \text{ and all } J \subseteq \{1, \ldots, N\} \Big\},$$

where N_i^k and N_j^k are the indices of the kth nearest neighbors of x_i and x_j, respectively.

2.3 Persistent homology

In this section, we define persistence modules, persistent homology (PH), and persistence diagrams. We assume the reader is familiar with homology. (A good introduction to homology and algebraic topology is [Hat02].) References for the rest of this section can be found in [ZC05, CdG16].

A *persistence module* $\mathbb{V}$ over $\mathbb{R}$ is a set $\{\mathbb{V}_r\}_{r\in\mathbb{R}}$ of vector spaces with a set $\{v_r^s : \mathbb{V}_r \to \mathbb{V}_s \text{ for all } r \leq s\}$ of linear maps that satisfy the composition law $v_s^t \circ v_r^s = v_r^t$ for all $r \leq s \leq t$. If $\{\mathcal{K}_r\}_{r\in\mathbb{R}}$ is a filtered complex, its *qth-persistent homology* over a field $\mathbb{F}$ is the persistence module $\{H_q(\mathcal{K}_r, \mathbb{F})\}_{r\in\mathbb{R}}$. The linear maps required for the persistence module are the maps $\imath_r^s : H_q(\mathcal{K}_r, \mathbb{F}) \to H_q(\mathcal{K}_s, \mathbb{F})$ induced by the inclusions $\mathcal{K}_r \hookrightarrow \mathcal{K}_s$ for all $r \leq s$. We will sometimes omit the field $\mathbb{F}$ from our notation when a fixed field has already been chosen.

As r increases, new homology classes are "born" and old homology classes "die". A homology class $\gamma \in H_q(\mathcal{K}_r, \mathbb{F})$ is *born* at filtration-parameter value b if it is not in the image

of ι_r^b for any $r < b$. The homology class γ subsequently *dies* at $d \geq b$ if $\iota_b^d(\gamma) = 0$ and $\iota_b^r(\gamma) \neq 0$ for all $r < d$. Not all homology classes die; if $\iota_b^r(\gamma) \neq 0$ for all $r > b$, then γ never dies.

For the remainder of this dissertation, we will always assume that $\mathcal{K} := \bigcup_{r \in \mathbb{R}} \mathcal{K}_r$ is finite—all simplicial complexes discussed thus far have been finite. Then there is a sequence $\{r_i\}$, with $r_i < r_{i+1}$, such that $\mathcal{K}_{r_i} = \mathcal{K}_r$ for all $r \in [r_i, r_{i+1})$. The set $\{r_i\}_i$ is the set of filtration values of the simplices in $\mathcal{K}$; these are the critical values at which new simplices are added. The simplicial complexes $\{\mathcal{K}_{r_i}\}$ form a nested sequence

$$\emptyset \subseteq \mathcal{K}_{r_1} \subseteq \mathcal{K}_{r_2} \subseteq \cdots \subseteq \mathcal{K} \tag{2.2}$$

of simplicial complexes; Equation 2.2 is an alternative definition of a filtered complex (see Definition 2.2.1 for comparison).

The Fundamental Theorem of Persistent Homology, stated below, shows that we can decompose the persistence module in a way that yields compatible sets of generators for each $H_q(\mathcal{K}_r, \mathbb{R})$. The direct sum $\bigoplus_i H_q(\mathcal{K}_{r_i}, \mathbb{F})$ has the structure of a graded module over the graded ring $\mathbb{F}[x]$. The action of x on a homogenous element $\gamma \in H_q(\mathcal{K}_{r_i}, \mathbb{F})$ is $x\gamma = \iota_{r_i}^{r_{i+1}}(\gamma)$.

Theorem 2.3.1 (Fundamental Theorem of Persistent Homology [ZC05])**.** *The graded $\mathbb{F}[x]$-module $\bigoplus_i H_q(\mathcal{K}_{r_i}, \mathbb{F})$ is isomorphic to*

$$\Big(\bigoplus_i \Sigma^{a_i}\mathbb{F}[x]\Big) \oplus \Big(\bigoplus_j \Sigma^{b_j}\mathbb{F}[x]/(x^{c_j})\Big) \tag{2.3}$$

for some integers $\{a_i\}$, $\{b_j\}$, $\{c_j\}$, where $\Sigma^m\mathbb{F}[x]$ denotes an m-shift upward in grading.

A $\Sigma^{a_i}\mathbb{F}[x]$ summand corresponds to a homology class that is born at filtration-parameter value r_{a_i} and never dies. A $\Sigma^{b_j}\mathbb{F}[x]/(x^{c_j})$ summand corresponds to a homology class that is born at filtration-parameter value r_{b_j} and dies at filtration-parameter value $r_{b_j} + r_{c_j}$.

The Fundamental Theorem of Persistent Homology allows us to define a persistence diagram, which is a multiset of points in the extended plane $\overline{\mathbb{R}}^2$ that summarizes persistent

homology. Given a decomposition in the form of Equation 2.3, the *qth-persistence diagram* consists of the points (r_{a_i}, ∞) for all i, the points $(r_{b_j}, r_{b_j} + r_{c_j})$ for all j, and all points on the diagonal (with infinite multiplicity). Note that off-diagonal points may also appear with multiplicity. The points on the diagonal are included for technical reasons; one can think of them as homology classes that die instantaneously. We denote the qth-persistence diagram of a filtered complex $(\mathcal{K}, f)$ by $\mathrm{PD}_q(f)$. More generally, persistence diagrams can be defined for any persistence module; we denote the persistence diagram of a persistence module $\mathbb{V}$ by $\mathrm{PD}(\mathbb{V})$.

A common way to compare two persistence diagrams is to compute their *bottleneck distance*. If $\mathbb{V}$ and $\mathbb{U}$ are two persistence modules, then the bottleneck between $\mathrm{PD}(\mathbb{V})$ and $\mathrm{PD}(\mathbb{U})$ is defined to be

$$W_\infty(\mathrm{PD}(\mathbb{V}), \mathrm{PD}(\mathbb{U})) := \inf_{\eta} \sup_{x \in \mathrm{PD}(\mathbb{V})} \left\| x - \eta(x) \right\|_\infty ,$$

where the infimum is taken over all bijections $\eta : \mathrm{PD}(\mathbb{V}) \to \mathrm{PD}(\mathbb{U})$.

2.4 Computing persistent homology and (birth, death) simplex pairs

For the remainder of Chapter 2, we compute persistent homology over the field $\mathbb{F} = \mathbb{Z}/2\mathbb{Z}$. Computing the persistent homology of a filtered complex $(\mathcal{K}, f)$ over $\mathbb{F} = \mathbb{Z}/2\mathbb{Z}$ can be reduced to computing the set of "birth" and "death" simplices for the generating homology classes. Informally, a birth simplex σ_b is a q-simplex that creates a new q-dimensional homology class when it is added to the filtered complex and a death simplex is a $(q+1)$-simplex that destroys a q-dimensional homology class when it is added to the filtered complex. For example, in Figure 2.1, the 1D PH has one generator. Its birth simplex is the 1-simplex $(0, 3)$ and its death simplex is the 2-simplex $(0, 2, 3)$. For every pair (σ_b, σ_d) of (birth, death) simplices, the persistence diagram contains the point $(f(\sigma_b), f(\sigma_d))$. For every unpaired birth

simplex σ_b, the persistence diagram contains the point $(f(\sigma_b), \infty)$. We will formally define birth and death simplices later in Section 2.4.

In [ELZ22], Edelsbrunner and Harer presented an algorithm for computing the (birth, death) simplex pairs of a filtration $f : \mathcal{K} \to \mathbb{R}$. Let $\sigma_1, \ldots, \sigma_N$ be the simplices of $\mathcal{K}$, indexed such that $i < j$ if σ_i is a proper face of σ_j.

Definition 2.4.1. The *simplex ordering* induced by f is the strict partial order $\prec_f$ on $\mathcal{K}$ such that $\sigma_i \prec_f \sigma_j$ if and only if $f(\sigma_i) < f(\sigma_j)$.

If the simplex orderings $\prec_{f_1}, \prec_{f_2}$ induced by filtrations f_1, f_2 (respectively) are the same, then $f_1(\sigma_i) < f_1(\sigma_j)$ if and only if $f_2(\sigma_i) < f_2(\sigma_j)$ and $f_1(\sigma_i) = f_1(\sigma_j)$ if and only if $f_2(\sigma_i) = f_2(\sigma_j)$.

The algorithm of [ELZ22] requires a *compatible simplex indexing*.

Definition 2.4.2. A *simplex indexing* is an injective function $\mathrm{idx} : \mathcal{K} \to \{1, \ldots, N\}$.

Definition 2.4.3. A *compatible simplex indexing* is a function $\mathrm{idx} : \mathcal{K} \to \{1, \ldots, N\}$ such that $\mathrm{idx}(\sigma_i) < \mathrm{idx}(\sigma_j)$ if $\sigma_i \prec_f \sigma_j$ or σ_i is a proper face of σ_j. Because a compatible simplex indexing may not be unique, we fix the *simplex indexing induced by* f to be the unique function $\mathrm{idx}_f : \mathcal{K} \to \{1, \ldots, N\}$ such that $\mathrm{idx}_f(\sigma_i) < \mathrm{idx}_f(\sigma_j)$ if either $\sigma_i \prec_f \sigma_j$ or if $f(\sigma_i) = f(\sigma_j)$ and $i < j$.

The function idx_f is a compatible simplex indexing because if σ_i is a proper face of σ_j, then $i < j$ and $f(\sigma_i) \leq f(\sigma_j)$. The sequence $\mathrm{idx}_f^{-1}(1), \ldots, \mathrm{idx}_f^{-1}(N)$ of simplices is ordered by the value of f on each simplex, with ties broken by the order of the simplices in the sequence $\sigma_1, \ldots, \sigma_N$. The indexing idx_f is defined such that if we define $\mathcal{K}'_i := \{\sigma \in \mathcal{K} \mid \mathrm{idx}_f(\sigma) \leq i\}$, then

$$\mathcal{K}'_1 \subseteq \mathcal{K}'_2 \subseteq \cdots \subseteq \mathcal{K}'_N$$

is a nested sequence of simplicial complexes and if $r_j = f(\sigma_{i_1}) = \cdots = f(\sigma_{i_k})$, where $i_1 < \cdots < i_k$ and $\{r_j\} = \mathrm{Im}(f)$, with $r_j < r_{j+1}$, then

$$\mathcal{K}_{r_j} = \mathcal{K}'_{i_1} \subset \mathcal{K}'_{i_2} \subset \cdots \subset \mathcal{K}'_{i_k} \subset \mathcal{K}_{r_{j+1}}\,.$$

In other words, $\{\mathcal{K}'_i\}$ is a refinement of $\{\mathcal{K}_{r_j}\}$.

Let D be the *boundary matrix* compatible with the simplex indexing $\text{idx}_f : \mathcal{K} \to \{1, \ldots, N\}$ that is induced by the filtration function f. More precisely, let D be the matrix whose (i, j)th entry is

$$D_{ij} = \begin{cases} 1\,, & \text{idx}_f^{-1}(i) \text{ is an } (m-1)\text{-dimensional face of the } m\text{-dimensional simplex } \text{idx}_f^{-1}(j) \\ 0\,, & \text{otherwise.} \end{cases}$$

We decompose the boundary matrix D into a matrix product $D = RU$ such that U is upper triangular and R is a binary matrix that is "reduced". A binary matrix R is *reduced* if $\text{low}_R(j) \neq \text{low}_R(j')$ whenever $j \neq j'$ are the indices of nonzero columns in R, where $\text{low}_R(\cdot)$ is the *pairing function*. The quantity $\text{low}_R(j)$ is the row index of the last 1 in column j if column j is nonzero and undefined if column j is the zero vector. An RU decomposition can be computed in $\mathcal{O}(N^3)$ time [ELZ22, EH10].

Cohen-Steiner et al. [CEM06] showed that the pairing function $\text{low}_R(j)$ depends only on the boundary matrix D and not on the particular reduced binary matrix R in the decomposition $D = RU$. A pair $(\text{idx}^{-1}(i), \text{idx}^{-1}(j))$ of simplices with $i = \text{low}_R(j)$ represents a persistent homology class. The *birth simplex* $\text{idx}^{-1}(i)$ creates the homology class and the *death simplex* $\text{idx}^{-1}(j)$ destroys the homology class. The two simplices in a pair have consecutive dimensions; that is, if $\dim(\text{idx}^{-1}(i)) = q$, then $\dim(\text{idx}^{-1}(j)) = q + 1$. If $\dim(\text{idx}^{-1}(i)) = q$ and $\dim(\text{idx}^{-1}(j)) = p + 1$, then a point with coordinates $(f(\text{idx}^{-1}(i), f(\text{idx}^{-1}(j)))$ is added to the qth persistence diagram. We refer to $f(\text{idx}^{-1}(i))$ as its *birth* and to $f(\text{idx}^{-1}(j))$ as its *death*. Some simplices are not paired. If $i \neq \text{low}_R(j)$ for all j, then the simplex $\text{idx}^{-1}(i)$ is a birth simplex for a homology class that never dies. Its birth is $f(\text{idx}^{-1}(i))$ and its death is ∞. If $\dim(\text{idx}^{-1}(i)) = q$, then a point with coordinates $(f(\text{idx}^{-1}(i)), \infty)$ is added to the qth persistence diagram.

The following lemma is a straightforward corollary of the work in [EH10], and we will rely on it repeatedly in Chapter 3.

Lemma 2.4.4 ([EH10]). If $f_0, f_1 : \mathcal{K} \to \mathbb{R}$ are two filtration functions such that $\prec_{f_0}$ is the same as $\prec_{f_1}$, then $\text{idx}_{f_0} = \text{idx}_{f_1}$ and f_1 and f_2 both induce the same set of (birth, death) simplex pairs.

2.5 Vineyards

Let $\mathcal{K}$ be a simplicial complex. A 1-*parameter filtration function* on $\mathcal{K}$ is a function $f : \mathcal{K} \times I \to \mathbb{R}$, where $I = [t_0, t_1]$ is an interval in $\mathbb{R}$, such that $f(\cdot, t)$ is a filtration function on $\mathcal{K}$ for all $t \in I$. For each $t \in I$, the r-sublevel sets $\mathcal{K}_r^t = \{\sigma \in \mathcal{K} \mid f(\sigma, t) \leq r\}$ form a filtered complex. The set $\{\{\mathcal{K}_r^t\}_{r\in\mathbb{R}}\}_{t\in I}$ is a set of filtrations parameterized by $t \in I$. For each $t \in I$, one can compute the persistence diagram $\text{PD}(f(\cdot, t))$. The associated *vineyard* is the 1-parameter set $\{\text{PD}(f(\cdot, t))\}_{t\in I}$ of persistence diagrams. We visualize the vineyard in $\overline{\mathbb{R}}^2 \times I$ as a continuous stack of PDs (see Figure 1.2). The points in the PDs trace out curves with time; these curves are the *vines*.

At any single point $t \in I$, a point in $\text{PD}(f(\cdot, t))$ corresponds to a homology class that has a (birth simplex, death simplex) pair. A vine in a vineyard corresponds to a sequence of (birth simplex, death simplex) pairs. For a given vine, we denote the pair at t by $(\sigma_b(t), \sigma_d(t))$. We define the *persistence* of the vine to be $\int_{t_0}^{t_1}[f(t, \sigma_d(t) - f(t, \sigma_b(t))]\text{d}t$.

An algorithm for computing vineyards is given by [CEM06], and we review it here. Define $\text{idx}_f : \mathcal{K} \times I \to \{1, \ldots, N\}$ by defining $\text{idx}(\cdot, t) := \text{idx}_{f(\cdot,t)}$ for all $t \in I$, where $\text{idx}_{f(\cdot,t)}$ is the simplex indexing induced by $f(\cdot, t)$ (see Definition 2.4.3). The simplex indexing only changes at times t_* at which $f(\sigma, t_*) = f(\tau, t_*)$ for some $\sigma, \tau \in \mathcal{K}$. At t_*, the simplex indexing of σ and τ may change. (If their indices change and they are the unique pair such that $f(\sigma, t_*) = f(\tau, t_*)$, then σ and τ are transposed in the simplex indexing and σ and τ have consecutive indices in the indexing. Otherwise, there is a sequence of such transpositions.) Let $D(t)$ denote the boundary matrix compatible with the simplex indexing $\text{idx}_f(\cdot, t)$ and let $\text{low}_R(\cdot, t)$ denote the corresponding pairing function. One computes $D(t_0)$ at the initial time

t_0 and an RU decomposition $D(t_0) = R(t_0)U(t_0)$. The initial pairing function $\mathrm{low}_R(\cdot, t_0)$ is read off from the initial RU decomposition. Then we sweep through the intersections t_* (from left to right) at which the simplex indexing changes. At each t_*, we update the simplex indexing, RU decomposition, and pairing function. (See Secion 2.6 for more details on how the pairing function is updated.) This updating procedure yields the birth and death simplices for the filtration function $f(\cdot, t_*)$, which one can then use to obtain the new persistence diagram.

The procedure above is an efficient way of computing the diagrams $\mathrm{PD}(f(\cdot, t))$ for all $t \in I$. At worst, updating $R(t_*)$ requires adding one column to another and adding one row to another, and similarly for $U(t_*)$. The addition of columns and rows is an $\mathcal{O}(N)$ operation, although in experiments, the authors of [CEM06] found that updating $R(t_*)$ and $U(t_*)$ can be done in approximately constant time if one uses the sparse matrix representations that are given in [CEM06]. If there is a single transposition of simplices at t_*, then at most two (birth, death) simplex pairs are updated, and these updates occur in $\mathcal{O}(1)$ time.

A special type of vineyard is a *piecewise-linear vineyard*. If we are only given $f(\sigma, t_i)$ at discrete time steps t_i, then for all i we extend $f(\sigma, t)$ to $t \in [t_i, t_{i+1}]$ by linear interpolation. In this case, one can compute the transposition times t_* by using the Bentley–Ottman planesweep algorithm [BCK08]. This is because computing when/if two simplices σ, τ get transposed in $[t_i, t_{i+1}]$ is equivalent to finding the intersection (if it exists) between the lines

$$y = \frac{f(\sigma, t_{i+1}) - f(\sigma, t_i)}{t_{i+1} - t_i}(t - t_i) + f(\sigma, t_i)\,,$$
$$y = \frac{f(\tau, t_{i+1}) - f(\tau, t_i)}{t_{i+1} - t_i}(t - t_i) + f(\tau, t_i)\,.$$

2.6 Updating PH when the simplex indexing is updated

One of the main contributions of [CEM06], in which vineyards were introduced, is an algorithm for updating the (birth, death) simplex pairs when the simplex indexing changes. We

review the relevant details in this subsection.

Suppose that $\mathrm{idx}_{f_0}, \mathrm{idx}_{f_1} : \mathcal{K} \to \{1, \ldots, N\}$ are the simplex indexings that are induced by filtrations $f_0, f_1 : \mathcal{K} \to \mathbb{R}$ (respectively), and suppose that idx_{f_0} and idx_{f_1} differ only by a transposition of a pair (σ, τ) of consecutive simplices. That is, $\mathrm{idx}_{f_0}(\tau) = \mathrm{idx}_{f_0}(\sigma) + 1$, $\mathrm{idx}_{f_1}(\tau) = \mathrm{idx}_{f_0}(\sigma)$, and $\mathrm{idx}_{f_1}(\sigma) = \mathrm{idx}_{f_0}(\tau)$. Let $S_{\mathrm{idx}_{f_0}}$ and $S_{\mathrm{idx}_{f_1}}$ be the sets of (birth, death) simplex pairs for f_0 and f_1, respectively.[2] The update rule of [CEM06] gives us a bijection $\phi^{\mathrm{idx}_{f_0}, \mathrm{idx}_{f_1}} : S_{\mathrm{idx}_{f_0}} \to S_{\mathrm{idx}_{f_1}}$.

We review the key properties of the bijection $\phi^{\mathrm{idx}_{f_0}, \mathrm{idx}_{f_1}}$. We write

$$\phi^{\mathrm{idx}_{f_0}, \mathrm{idx}_{f_1}} = (\phi_b^{\mathrm{idx}_{f_0}, \mathrm{idx}_{f_1}}, \phi_d^{\mathrm{idx}_{f_0}, \mathrm{idx}_{f_1}}),$$

where $\phi_b^{\mathrm{idx}_{f_0}, \mathrm{idx}_{f_1}} : S_{\mathrm{idx}_{f_0}} \to \mathcal{K}$ maps a simplex pair $(\sigma_b, \sigma_d) \in S_{\mathrm{idx}_{f_0}}$ to the birth simplex of $\phi^{\mathrm{idx}_{f_0}, \mathrm{idx}_{f_1}}((\sigma_b, \sigma_d))$ and $\phi_d^{\mathrm{idx}_{f_0}, \mathrm{idx}_{f_1}} : S_{\mathrm{idx}_{f_0}} \to \mathcal{K}$ maps $(\sigma_b, \sigma_d) \in S_{\mathrm{idx}_{f_0}}$ to the death simplex of $\phi^{\mathrm{idx}_{f_0}, \mathrm{idx}_{f_1}}((\sigma_b, \sigma_d))$. If $(\sigma_b, \sigma_d) \in S_{\mathrm{idx}_{f_0}}$ is a pair such that $\sigma_b, \sigma_d \notin \{\sigma, \tau\}$, then $\phi^{\mathrm{idx}_{f_0}, \mathrm{idx}_{f_1}}((\sigma_b, \sigma_d)) = (\sigma_b, \sigma_d)$. If $(\sigma_b^1, \sigma_d^1) \in S_{\mathrm{idx}_{f_0}}$ is the pair that contains σ, then let $\lambda \in \{b, d\}$ be the index such that $\sigma_\lambda^1 = \sigma$. Similarly, if $(\sigma_b^2, \sigma_d^2) \in S_{\mathrm{idx}_{f_0}}$ is the pair that contains τ, then let $\mu \in \{b, d\}$ be the index such that $\sigma_\mu^1 = \tau$. The key fact about the update rule of [CEM06] is that $\phi^{\mathrm{idx}_{f_0}, \mathrm{idx}_{f_1}}$ is defined such that either

$$\phi^{\mathrm{idx}_{f_0}, \mathrm{idx}_{f_1}}((\sigma_b^1, \sigma_d^2)) = (\sigma_b^1, \sigma_d^1),$$
$$\phi^{\mathrm{idx}_{f_0}, \mathrm{idx}_{f_1}}((\sigma_b^2, \sigma_d^2)) = (\sigma_b^2, \sigma_d^2),$$

or

$$\phi_\lambda^{\mathrm{idx}_{f_0}, \mathrm{idx}_{f_1}}((\sigma_b^1, \sigma_d^1)) = \tau, \qquad \phi_{\lambda^c}^{\mathrm{idx}_{f_0}, \mathrm{idx}_{f_1}}((\sigma_b^1, \sigma_d^1)) = \sigma_{\lambda^c}^1,$$
$$\phi_\mu^{\mathrm{idx}_{f_0}, \mathrm{idx}_{f_1}}((\sigma_b^2, \sigma_d^2)) = \sigma, \qquad \phi_{\mu^c}^{\mathrm{idx}_{f_0}, \mathrm{idx}_{f_1}}((\sigma_b^2, \sigma_d^2)) = \sigma_{\mu^c}^1,$$

[2] Recall that, by Lemma 2.4.4, the pairs depend only on the simplex orderings idx_{f_0}, idx_{f_1}, which is why we label the sets by their associated simplex indexings.

where

$$\lambda^c := \begin{cases} b\,, & \lambda = d \\ d\,, & \lambda = b \end{cases} \qquad \mu^c := \begin{cases} b\,, & \mu = d \\ d\,, & \mu = b\,. \end{cases}$$

In other words, either $\phi^{\mathrm{idx}_{f_0}, \mathrm{idx}_{f_1}}$ is the identity map or $\phi^{\mathrm{idx}_{f_0}, \mathrm{idx}_{f_1}}$ swaps σ and τ in the pairs that contain them. The particular case depends on the order of $f_0(\sigma_b^1), f_0(\sigma_d^1), f_0(\sigma_b^2), f_0(\sigma_d^2)$ (see [CEM06] for details; they are not relevant to the present thesis).

More generally, suppose that idx_{f_0}, idx_{f_1} are the simplex indexings induced by any two filtrations f_0, f_1, where idx_{f_0} and idx_{f_1} are no longer required to differ only by the transposition of two consecutive simplices. Let $S_{\mathrm{idx}_{f_0}}$ and $S_{\mathrm{idx}_{f_1}}$ be the sets of (birth, death) simplex pairs for f_0 and f_1, respectively. The update rule of Cohen-Steiner et al. [CEM06] defines a bijection $\phi^{\mathrm{idx}_{f_0}, \mathrm{idx}_{f_1}} : S_{\mathrm{idx}_{f_0}} \to S_{\mathrm{idx}_{f_1}}$ as follows. Every permutation can be decomposed into a sequence of transpositions that transpose consecutive elements, so there is a sequence $\zeta_0, \ldots, \zeta_m$ of simplex indexings such that $\zeta_0 = \mathrm{idx}_{f_0}$, $\zeta_m = \mathrm{idx}_{f_1}$, and ζ_i, ζ_{i+1} differ only by the transposition of two consecutive simplices. Cohen-Steiner et al. [CEM06] defined

$$\phi^{\mathrm{idx}_{f_0}, \mathrm{idx}_{f_1}} := \phi^{\zeta_{m-1}, \zeta_m} \circ \cdots \circ \phi^{\zeta_0, \zeta_1}\,. \tag{2.4}$$

Remark 2.6.1. If idx_{f_0}, idx_{f_1} do not differ by only the transposition of two consecutive simplices, then the sequence $\zeta_0, \ldots, \zeta_m$ is not unique. Unfortunately, the definition of $\phi^{\mathrm{idx}_{f_0}, \mathrm{idx}_{f_1}}$ does depend on the sequence $\zeta_0, \ldots, \zeta_m$ in its definition. This is implicitly shown in Prop. 3.4.3.

2.7 Cellular sheaves

We review the definition of a cellular sheaf; references include [HG19, Cur14]. A *cell complex* is a topological space Y with a partition into a set $\{Y_\alpha\}_{\alpha \in P_Y}$ of subspaces (the *cells* of the cell complex) that satisfy the following conditions:

1. Every cell Y_α is homeomorphic to $\mathbb{R}^{k_\alpha}$ for some $k_\alpha \geq 0$. The cell Y_α is a k_α-*cell*.

2. For every cell Y_α, there is a homeomorphism $\phi_\alpha : B^{k_\alpha} \to \overline{X_\alpha}$, where B^{k_α} is the closed k_α-dimensional ball, such that $\phi_\alpha(\text{int}(B^{k_\alpha})) = X_\alpha$.

3. **Axiom of the frontier:** If the intersection $\overline{Y_\beta} \cap Y_\alpha$ is nonempty, then $Y_\alpha \subseteq \overline{Y_\beta}$. We say that Y_α is a *face* of Y_β.

4. **Locally finite:** Every $x \in X$ has an open neighborhood U such that U intersects finitely many cells.

For example, a polyhedron is a cell complex whose k-cells are the k-dimensional faces of the polyhedron. A graph is a another example of a cell complex; the 0-cells are the vertices and the 1-cells are the edges.

We will first review the most general definition of cellular sheaves, which uses category theory, and then we will specialize to the case of interest for this dissertation, which does not require category theory. Let Y be a cell complex with cells $\{Y_\alpha\}_{\alpha \in P_Y}$, and let $\mathcal{D}$ be a category. The set P_Y is a poset with the relation $\alpha \leq \beta$ if $Y_\alpha \subseteq \overline{Y_\beta}$. A *$\mathcal{D}$-valued cellular sheaf on Y* consists of the following:

1. An assignment of an object $\mathcal{F}(Y_\alpha) \in \mathcal{D}$ (the *stalk* of $\mathcal{F}$ at Y_α) for every cell Y_α in Y.

2. A morphism $\mathcal{F}_{\alpha \leq \beta} : F(Y_\alpha) \to F(Y_\beta)$ (a *restriction map*) whenever Y_α is a face of Y_β. The morphisms must satisfy the *composition condition*:

$$\mathcal{F}_{\beta \leq \gamma} \circ \mathcal{F}_{\alpha \leq \beta} = \mathcal{F}_{\alpha \leq \gamma} \tag{2.5}$$

   whenever $\alpha \leq \beta \leq \gamma$.

Equivalently, a $\mathcal{D}$-valued cellular sheaf on Y is a functor $\mathcal{F} : P_Y \to \mathcal{D}$, where P_Y is considered as a category.

A *global section* of a $\mathcal{D}$-valued cellular sheaf is a function

$$s : \{Y_\alpha\}_{\alpha \in P_Y} \to \bigcup_\alpha F(Y_\alpha)$$

such that

1. $s(Y_\alpha) \in \mathcal{F}(Y_\alpha)$ (i.e., $s(Y_\alpha)$ is a choice of element in the stalk at Y_α) and
2. if $\alpha \leq \beta$, then $s(Y_\beta) = \mathcal{F}_{\alpha \leq \beta}(s(Y_\alpha))$.

In Chapter 3, we will consider a **Set**-valued cellular sheaf. The objects of the category **Set** are sets, and the morphisms between sets A and B are the functions from A to B. A **Set**-valued cellular sheaf on a cell complex Y consists of the following:

1. A set $\mathcal{F}(Y_\alpha)$ for every cell Y_α in Y.
2. A function $\mathcal{F}_{\alpha \leq \beta} : \mathcal{F}(Y_\alpha) \to \mathcal{F}(Y_\beta)$ whenever Y_α is a face of Y_β. The functions must satisfy the condition
$$\mathcal{F}_{\beta \leq \gamma} \circ \mathcal{F}_{\alpha \leq \beta} = \mathcal{F}_{\alpha \leq \gamma}$$
whenever $\alpha \leq \beta \leq \gamma$.

2.8 Riemannian geometry

We briefly review the necessary background from Riemannian geometry. For further reading, I recommend a textbook such as [Pet06]. A *Riemannian manifold* (M, g) is a smooth manifold M with a Riemannian metric g, which defines a smoothly-varying inner product on each tangent space T_xM. More formally, a *Riemannian metric* is an assignment of a symmetric, bilinear, positive-definite map g_x on the tangent space T_xM for each $x \in M$. For example, in Euclidean space, the *canonical Euclidean metric* $\bar{g}$ assigns the usual Euclidean inner product to each tangent space.

All manifolds can be given a Riemannian metric. To see this, recall that all manifolds can be embedded into Euclidean space. If $F : (M, g) \to (M', g')$ is any smooth map between two Riemannian manifolds, the *pullback* of g' is the Riemannian metric F^*g' that is defined by $(F^*g')_x(v, w) := g'_{F(x)}(dF_x v, dF_x w)$ for all $x \in M$ and all $v, w \in T_xM$. Therefore, if

$\iota : M \hookrightarrow \mathbb{R}^m$ is an embedding, the canonical Euclidean metric $\bar{g}$ pulls back to a Riemmanian metric $\iota^*\bar{g}$ on M. For example, if M is a submanifold of $\mathbb{R}^m$ and $\iota : M \hookrightarrow \mathbb{R}^M$ is the inclusion, then $\iota^*\bar{g}_x$ is the restriction of $\bar{g}_x$ to T_xM. We refer to $\iota^*\bar{g}$ as the *Euclidean-induced Riemannian metric*.

A Riemannian metric induces a metric structure. The *norm* of a vector $v \in T_xM$ is $\|v\| := g_x(v,v)^{1/2}$. The *length* of a continuously differentiable path $\gamma : [a,b] \to M$ is $L(\gamma) := \int_a^b \left\|\gamma'(t)\right\| \mathrm{d}t$. The *geodesic distance* between two points x and y in the same connected component of M is

$$d_{M,g}(x,y) := \inf\{L(\gamma) \mid \gamma : [a,b] \to M \text{ is a } C^1 \text{ path such that } \gamma(a) = x \text{ and } \gamma(b) = y\}.$$

If (M,g) is *complete*, then the infimum is achieved by a *geodesic*, which is a curve that locally minimizes length. If x and y are in different connected components, then the geodesic distance between them is infinite. The closed *geodesic ball* of radius r at point x is

$$B^M(x,r) := \{y \in M \mid d_{M,g}(x,y) \leq r\}.$$

One can define a probability measure on a Riemannian manifold as follows. A Riemannian metric induces a *volume form* dV, which is the unique n-form on M that equals 1 on all positively oriented orthonormal bases. In local coordinates, the expression for the volume form is

$$dV = \sqrt{|g|} dx^1 \wedge \cdots \wedge dx^n.$$

The volume form induces a *Riemannian measure* μ on M. The measure of a Borel set $A \subseteq M$ is $\mu(A) = \int_A dV$, and the volume of M is $\mu(M)$. Given a smooth probability density function $\rho : M \to (0,\infty)$, the induced probability measure is

$$\mathbb{P}[A] := \int_A \rho dV$$

for Borel sets $A \subseteq M$. A good reference for probability and statistics on Riemannian manifolds is [Pen06].

In Chapter 7, we will consider the *scalar curvature* of a Riemannian manifold. Scalar curvature characterizes the rate at which the volume of a geodesic ball $B^M(x,r)$ grows as r grows. As $r \to 0$, the scalar curvature $S(x)$ at $x \in M$ has the following relationship to geodesic ball volume:

$$\frac{\text{vol}(B^M(x,r))}{v_n r^n} = 1 - \frac{S(x)}{6(n+2)} r^2 + \mathcal{O}(r^4)\,, \tag{2.6}$$

where n is the dimension of the manifold, v_n is the volume of a unit Euclidean n-ball, and $v_n r^n$ is the volume of a Euclidean n-ball of radius r. For example, if $S(x)$ is negative (respectively, positive), then the volume of a small geodesic ball that is centered at x tends to be larger (respectively, smaller) than the volume of an n-dimensional Euclidean ball of the same radius.

CHAPTER 3

Persistence-Diagram Bundles: A Multidimensional Generalization of Vineyards

This chapter is adapted from [Hic22c].

3.1 Introduction

Developing new methods for analyzing how the topology of a data set changes as multiple parameters vary is an active area of research [BL22]. For example, if a point cloud evolves over time (i.e., it is a dynamic metric space), then one may be interested in using time as a second parameter, in addition to the filtration parameter r. Common examples of time-evolving point clouds include swarming or flocking animals whose positions and/or velocities are represented by points ([CJ17, XAT22, M21]). In such cases, one can obtain a filtered complex $\mathcal{K}^t_{r_0} \subseteq \mathcal{K}^t_{r_1} \subseteq \cdots \subseteq K^t_{r_n}$ at every time t by constructing, e.g., the Vietoris–Rips filtered complex for the point cloud at time t. It is also common to use the density of the point cloud as a parameter ([SCD22, CCM11, BMT17]). Many other parameters can also vary in the topological analysis of point clouds or other types of data sets.

One can use a vineyard [CEM06] to study a 1-parameter family of filtrations $\{\mathcal{K}^t_{r_0} \subseteq \mathcal{K}^t_{r_1} \subseteq \cdots \mathcal{K}^t_{r_n}\}_{t\in\mathbb{R}}$ such as that obtained from a time-varying point cloud. However, one cannot use a vineyard for a set of filtrations that is parameterized by a space that is not a subset of $\mathbb{R}$. For example, suppose that we have a time-varying point cloud whose dynamics depend on some system-parameter values $\mu_1, \ldots, \mu_m \in \mathbb{R}$. Many such systems exist. For

example, the D'Orsogna model is a multi-agent dynamical system that models attractive and repulsive interactions between particles [CDM07]. Each particle is represented by a point in a point cloud. In certain parameter regimes, there are interesting topological features, such as mills or double mills [TZH15a]. For each time $t \in \mathbb{R}$ and for each $\mu_1, \ldots, \mu_m \in \mathbb{R}$, one can obtain a filtered complex

$$\mathcal{K}_{r_0}^{t,\mu_1,\ldots,\mu_m} \subseteq \mathcal{K}_{r_1}^{t,\mu_1,\ldots,\mu_m} \subseteq \cdots \subseteq \mathcal{K}_{r_n}^{t,\mu_1,\ldots,\mu_m} \tag{3.1}$$

by constructing, e.g., the Vietoris–Rips filtered complex for the point cloud at time t at system-parameter values $\mu_1, \ldots, \mu_m$. A parameterized set of filtered complexes like the one in Equation 3.1 cannot be studied using a vineyard for the simple reason that there are too many parameters.

Such a parameterized set of filtered complexes also cannot be studied using multiparameter PH [CZ07], either. A *multifiltration* is a set $\{\mathcal{K}_{\boldsymbol{u}}\}_{\boldsymbol{u}\in\mathbb{R}^n}$ of simplicial complexes such that $\mathcal{K}_{\boldsymbol{u}} \subseteq \mathcal{K}_{\boldsymbol{v}}$ whenever $\boldsymbol{u} \leq \boldsymbol{v}$. *Multiparameter PH* is the $\mathbb{F}[x_1, \ldots, x_n]$ module obtained by applying homology to a multifiltration. (For more details, see reference [CZ07].) The parameterized set of filtered complexes in Equation 3.1 is not typically a multifiltration because it is not necessarily the case that $\mathcal{K}_{r_i}^{t,\mu_1,\ldots,\mu_m} \not\subseteq \mathcal{K}_{r_i}^{t',\mu_1',\ldots,\mu_m'}$ for all values of t, t', $\{\mu_i\}$, and $\{\mu_i'\}$. Neither do we have any canonical simplicial map $\mathcal{K}_{r_i}^{t,\mu_1,\ldots,\mu_m} \to \mathcal{K}_{r_i}^{t',\mu_1',\ldots,\mu_m'}$. Therefore, we cannot use multiparameter PH.

3.1.1 Contributions

I introduce the concept of a *persistence-diagram (PD) bundle*, in which PH varies over an arbitrary "base space" B. A PD bundle gives a way of studying a *fibered filtration function*, which is a set $\{f_p : \mathcal{K}^p \to \mathbb{F}\}_{p\in B}$ of functions such that f_p is a filtration of a simplicial complex $\mathcal{K}^p$. At each $p \in B$, the sublevel sets of f_p form a filtered complex. For example, in Equation 3.1, we have $B = \mathbb{R}^{n+1}$ and we obtain a fibered filtration function $\{f_{t,\mu_1,\ldots,\mu_m} : \mathcal{K} \to \mathbb{R}\}_{(t,\mu_1,\ldots,\mu_m)\in\mathbb{R}^{m+1}}$ by defining $f_{t,\mu_1,\ldots,\mu_m}$ to be the filtration function associated with

the filtered complex in Equation 3.1. The associated PD bundle is the space of persistence diagrams $\mathrm{PD}(f_p)$ as they vary with $p \in B$ (see Definition 3.2.2). In the special case in which B is an interval in $\mathbb{R}$, a PD bundle is equivalent to a vineyard.

I prove that for "generic" fibered filtration functions (see Section 3.3.2), the base space B can be stratified in a way that makes PD bundles tractable to compute and analyze. Theorem 3.3.15 says that for a "generic" fibered filtration function on a smooth compact manifold B, the base space B is stratified such that within each stratum, there is a single PD "template" that can be used to obtain $PD(f_p)$ at any point p in the stratum. Proposition 3.3.5 shows that *all* "piecewise-linear" PD bundles (see Definition 3.2.4) have such a stratification. The template is a list of (birth, death) simplex pairs, and the diagram $\mathrm{PD}(f_p)$ is obtained by evaluating f_p on each simplex. In particular, when B is a smooth compact manifold, the number of strata is finite, so the PD bundle is determined by the PH at a finite number of points in the base space.

I show that unlike vineyards, PD bundles typically do not have individual "vines" because nontrivial global sections are not guaranteed to exist (see Proposition 3.4.3). In other words, given a point $z_0 \in PD(f_{p_0})$ for some $p_0 \in B$, it may not be possible to extend $p_0 \mapsto z_0$ to a continuous map $s : B \to E := \{(p, z) \mid z \in PD(f_p)\}$. This is a feature that gives PD bundles a richer mathematical structure than vineyards.

For any fibered filtration with a stratification as described above (see Theorem 3.3.15 and Proposition 3.3.5), I construct a "compatible cellular sheaf" (see Section 3.5.2) that stores the data in the PD bundle. Rather than analyzing the entire PD bundle, which consists of continuously varying PDs over the base space B, we can analyze the cellular sheaf, which is discrete. For example, in Proposition 3.5.4, I prove that an extension of $p_0 \mapsto z_0$ to a global section exists if a certain associated global section of the cellular sheaf exists. A compatible cellular sheaf stores sufficient data to reconstruct the associated PD bundle and analyze its sections.

I also give a simple example of vineyard instability in Appendix 3.A. It is often quoted

in the research literature that "vineyards are unstable"; however, this "well-known fact" has been shared only in private correspondence and, to the best of my knowledge, has never been published. The example of vineyard instability is furnished from an example in Proposition 3.4.3.

3.1.2 Related work

PD bundles are a generalization of vineyards, which were introduced in [CEM06]. Two other important special cases of PD bundles are the fibered barcode of a multiparameter persistence module [And13] and the persistent homology transform ($B = S^n$) from shape analysis [AEH06, TMB14]. I discuss the special case of fibered barcodes in detail in Section 3.2.2.2; the base space B is a subset of the space of lines in $\mathbb{R}^n$. The persistent homology transform (PHT) is defined for a constructible set $M \subseteq \mathbb{R}^{n+1}$. For any unit vector $v \in S^n$, one defines the filtration $M_r^v = \{x \in M \mid x \cdot v \leq r\}$ (i.e., the sublevel filtration of the height function with respect to the direction v). PHT is the map that sends $v \in S^n$ to the persistence diagram for the filtration $\{M_r^v\}_{r\in\mathbb{R}}$. The significance of PHT is that it is a sufficient statistic for shapes in $\mathbb{R}^2$ and $\mathbb{R}^3$ [TMB14]. Applications of PHT are numerous and include protein docking [WAB05], barley-seed shape analysis [AQO22], and heel-bone analysis in primates [TMB14].

For PHT, Curry et al. [CMT22] proved that the base space S^n is stratified such that the PHT of a shape M is determined by the PH of $\{M_r^v\}_{r\in\mathbb{R}}$ for finitely many directions $v \in S^n$ (one direction v per stratum). This is related to the stratification given by Theorem 3.3.15, in which I show that a "generic" PD bundle whose base space B is a compact smooth manifold (such as S^n) is similarly stratified and thus determined by finitely many points in B (one $p \in B$ per stratum). The primary difference between the stratifications in [CMT22] and Theorem 3.3.15 is that in [CMT22], each stratum is a subset in which the order of the vertices of a triangulated shape M (as ordered by the height function) is constant, whereas in Theorem 3.3.15, each stratum is a subset in which the order of the simplices (as ordered

by the filtration function) is constant.

The stratification that we study in the present chapter is used in [Hic22a] to develop an algorithm for computing "piecewise-linear" PD bundles (see Definition 3.2.4). The algorithm relies on the fact that for any piecewise-linear PD bundle on a compact triangulated base space B, there are a finite number of strata, so the PD bundle is determined by the PH at a finite number of points in B.

The existence (or nonexistence) of nontrivial global sections in PD bundles is related to the study of "monodromy" in fibered barcodes of multiparameter persistence modules [CEF13]. Cerri et al. [CEF13] constructed an example in which there is a path through the fibered barcode that loops around a "singularity" (a PD in the fibered barcode for which there is a point in the PD with multiplicity greater than 1) and finishes in a different place than where it starts.

3.1.3 Organization

This chapter proceeds as follows. In Section 3.2, we give the definition of a PD bundle, with some examples, and we compare PD bundles to multiparameter PH. In Section 3.3, we show how to stratify the base space B into strata in which the (birth, death) simplex pairs are constant (see Theorem 3.3.15 and Proposition 3.3.5). We discuss sections of PD bundles and the existence of monodromy in Section 3.4. We construct a compatible cellular sheaf in Section 3.5.2. We conclude and discuss possible directions for future research in Section 5.6. In Appendix 3.A, we use the example of monodromy from Section 3.4 to construct an example of vineyard instability. In Appendix 3.B, we provide technical details that are needed to prove Theorem 3.3.15.

3.2 Definition of a Persistence-Diagram Bundle

A vineyard is a 1-parameter set of persistence diagrams that is computed from a 1-parameter set of filtration functions on a simplicial complex $\mathcal{K}$. We generalize a vineyard to a "persistence-diagram bundle" as follows.

Definition 3.2.1 (Fibered filtration function)**.** A *fibered filtration function* is a set $\{f_p : \mathcal{K}^p \to \mathbb{R}\}_{p\in B}$, where B is a topological space, $\{\mathcal{K}^p\}_{p\in B}$ is a set of simplicial complexes parameterized by B, and f_p is a filtration function on $\mathcal{K}^p$.

When $\mathcal{K}^p \equiv \mathcal{K}$ for all $p \in B$, we define $f : \mathcal{K} \times B \to \mathbb{R}$ to be the function $f(\sigma, p) := f_p(\sigma)$. In a slight abuse of notation, we refer to $f : \mathcal{K} \times B \to \mathbb{R}$, rather than to $\{f_p : \mathcal{K} \to B\}_{p\in B}$, as the fibered filtration function. For all $p \in B$, the function $f(\cdot, p) : \mathcal{K} \to \mathbb{R}$ is a filtration of $\mathcal{K}$. For several examples with $\mathcal{K}^p \equiv \mathcal{K}$, see Section 3.2.1.

Definition 3.2.2 (Persistence-diagram bundle)**.** Let $\{f_p : \mathcal{K}^p \to \mathbb{R}\}_{p\in B}$ be a fibered filtration function. The *base* of the bundle is B. The *qth total space* of the bundle is $E := \{(p, z)\} \mid p \in B, z \in PD_q(f_p)\}$, with the subspace topology inherited from the inclusion $E \hookrightarrow B \times \overline{\mathbb{R}}^2$.[1] The *qth persistence-diagram bundle* is the triple (E, B, π), where $\pi : E \to B$ is the projection $(p, z) \mapsto p$.

For example, when B is an interval in $\mathbb{R}$ and $\mathcal{K}^p \equiv \mathcal{K}$, Definition 3.2.2 reduces to that of a vineyard: a 1-parameter set of PDs for a 1-parameter set of filtrations of $\mathcal{K}$. As discussed in Section 4.1, PHT is a special case with $B = S^d$. The fibered barcode of a multiparameter persistence module is another special case; we will discuss it in Section 3.2.2.2.

Remark 3.2.3. In Definition 3.2.2, we are suggestively using the language of fiber bundles. However, it is important to note that a PD bundle is not guaranteed to be a true fiber bundle.

[1]Technically, E is a multiset because persistence diagrams are multisets. However, when considering E as a topological space (which we do in Section 3.4 to study continuous paths in E and sections of the PD bundle), we consider E as a set.

The fibers need not be homeomorphic to each other for all $p \in B$. At "singularities" (points $p_* \in B$ at which $\mathrm{PD}(f_{p_*})$ has an off-diagonal point with multiplicity), points in $\mathrm{PD}(f_p)$ for nearby p may merge into each other, changing the homotopy type of the fiber. However, if $f : \mathcal{K} \times B \to \mathbb{R}$ is continuous and $p \in B$ is not a singularity, then there is a neighborhood $U \subseteq B$ and a homeomorphism $\phi : \pi^{-1}(U) \to U \times \mathrm{PD}(f_{p_*})$ that preserves fibers (i.e., a local trivialization).

As a special case of fibered filtration functions, we define *piecewise-linear fibered filtration functions*, which are simpler to analyze.

Definition 3.2.4 (Piecewise-linear fibered filtration function)**.** Let $\{f_p : \mathcal{K}^p \to \mathbb{R}\}_{p \in B}$ be a fibered filtration function such that $\mathcal{K}^p \equiv \mathcal{K}$. We define $f(\sigma, p) \coloneqq f_p(\sigma)$ for all $\sigma \in \mathcal{K}$ and $p \in B$. If B is a simplicial complex and $f(\sigma, \cdot)$ is linear on each simplex of B for all simplices $\sigma \in \mathcal{K}$, then f is a *piecewise-linear fibered filtration function.* The resulting PD bundle is a *piecewise-linear PD bundle.*

For instance, the fibered filtration function in Example 3.2.6, below, is piecewise linear.

3.2.1 Examples

The following are concrete examples of PD bundles. We begin with the example that motivated PD bundles in Section 4.1.

Example 3.2.5. Suppose that $X(t, \boldsymbol{\mu}) = \{x_1(t, \boldsymbol{\mu}), \ldots, x_k(t, \boldsymbol{\mu})\}$ is a point cloud that varies continuously with time $t \in \mathbb{R}$ and system-parameter values $\mu_1, \ldots, \mu_m \in \mathbb{R}$. We obtain a fibered filtration function $f : \mathcal{K} \times \mathbb{R}^{m+1} \to \mathbb{R}$ by defining $f(\cdot, (t, \boldsymbol{\mu})) : \mathcal{K} \to \mathbb{R}$ to be the Vietoris–Rips filtration function for the point cloud $X(t, \boldsymbol{\mu})$ at all $(t, \boldsymbol{\mu}) \in \mathbb{R}^{m+1}$ (or any other filtration for the point cloud at each $(t, \boldsymbol{\mu})$). The simplicial complex $\mathcal{K}$ is the simplicial complex that has a simplex for every subset of points in the point cloud.

Example 3.2.6. Consider a color image. Enumerate the pixels and let $r(i)$, $g(i)$, and $b(i)$ denote the red, green, and blue values of the ith pixel. Triangulate each pixel to obtain a

simplicial complex $\mathcal{K}$. (Every pixel is split into two triangles.) Let $B = \{(w_1, w_2) \in [0,1]^2 \mid 0 \leq w_1 + w_2 \leq 1\}$. For all $(w_1, w_2) \in B$, define $p(i, (w_1, w_2)) = w_1 r(i) + w_2 g(i) + (1 - w_1 - w_2) b(i)$. The function $p(i, (w_1, w_2))$ is a weighted average of the red, green, and blue values of the ith pixel. Define a piecewise-linear fibered filtration function $f : \mathcal{K} \times B \to \mathbb{R}$ as follows. For a 2-simplex σ, define $f(\sigma, \mathbf{w}) = p(i(\sigma), \mathbf{w})$, where $i(\sigma)$ is the pixel containing σ. For any other simplex $\sigma \in \mathcal{K}$, define $f(\sigma, \mathbf{w}) = \min\{f(\tau, \mathbf{w}) \mid \sigma \subseteq \tau, \dim(\tau) = 2\}$. At $\mathbf{w} = (1, 0)$, $\mathbf{w} = (0, 1)$, and $\mathbf{w} = (1, 1)$, the filtration function $f(\cdot, \mathbf{w})$ is the sublevel filtration by red, green, and blue pixel values, respectively. At all other $\mathbf{w} \in B$, the filtration function $f(\cdot, \mathbf{w})$ is the sublevel filtration by a weighted average of the red, green, and blue pixel values.

Example 3.2.7. Let $\mu_1, \ldots, \mu_m \in \mathbb{R}$ denote the system-parameter values of some discrete dynamical system. For given system-parameter values $\boldsymbol{\mu} \in \mathbb{R}^m$, let $x_i^{\boldsymbol{\mu}} \in \mathbb{R}^n$ be the solution at the ith time step and let $X(\boldsymbol{\mu}) = \{x_0^{\boldsymbol{\mu}}, \ldots, x_k^{\boldsymbol{\mu}}\}$ be the set of points obtained after the first k time steps. For example, persistent homology has been used to study orbits of the linked twist map (a discrete dynamical system [AEK17]). We obtain a fibered filtration function $f : \mathcal{K} \times \mathbb{R}^m \to \mathbb{R}$ by defining $f(\cdot, \boldsymbol{\mu}) : \mathcal{K} \to \mathbb{R}$ to be the Vietoris–Rips filtration function for the point cloud $X(\boldsymbol{\mu})$ (or any other filtration for the point cloud at each $\boldsymbol{\mu}$). The simplicial complex $\mathcal{K}$ has a simplex for every subset of points in the point cloud.

Example 3.2.8. Suppose that $X(t) = \{x_1(t), \ldots, x_k(t)\}$ is a time-varying point cloud in a compact triangulable subset $S \subseteq \mathbb{R}^n$. Let $\rho_h(\cdot, t)$ be a kernel density estimator at time t, with bandwidth parameter $h > 0$. For fixed h and t, we define a filtered complex by considering sublevel sets of ρ_h as follows. Let $\mathcal{K}$ be a triangulation of $S \subseteq \mathbb{R}^n$. A vertex v of $\mathcal{K}$ is included in the simplicial complex $\mathcal{K}_r^{(t,h)}$ if $\rho_h(v, t) \leq r$, and a simplex of $\mathcal{K}$ is included in $\mathcal{K}_r^{t,h}$ if all of its vertices are in $\mathcal{K}_r^{(t,h)}$. For each t and h, the set $\{\mathcal{K}_r^{(t,h)}\}_{r \in \mathbb{R}}$ is a filtered complex. We obtain a fibered filtration function $f : \mathcal{K} \times \mathbb{R}_+^2 \to \mathbb{R}$ by defining $f(\cdot, (t, h))$ to be the filtration function associated with the filtered complex $\{\mathcal{K}_r^{(t,h)}\}_{r \in \mathbb{R}}$.

Density sublevels of time-varying point clouds were also considered by Corcoran et al. in [CJ17], who studied a school of fish swimming in a shallow pool that was modeled as a

subset of $\mathbb{R}^2$. However, Corcoran et al. [CJ17] fixed a bandwidth parameter h and a sublevel r, and only studied how the PH changed with time (by using zigzag PH).

3.2.2 Comparison to multiparameter PH

Multiparameter PH was introduced in [CZ07]; see [BL22] for a review. Typically, a fibered filtration function does not induce a multifiltration, but the fibered barcode of a multiparameter peristence module is an example of a PD bundle.

3.2.2.1 Multifiltrations

We review the definition of a multifiltration and compare it to the definition of a fibered filtration.

Definition 3.2.9. A *multifiltration* is a set $\{\mathcal{K}_{\boldsymbol{u}}\}_{\boldsymbol{u}\in\mathbb{Z}^n}$ of simplicial complexes such that if $\boldsymbol{u} \leq \boldsymbol{v}$, then $\mathcal{K}_{\boldsymbol{u}} \subseteq \mathcal{K}_{\boldsymbol{v}}$.

The inclusion $\iota^{\boldsymbol{u},\boldsymbol{v}} : \mathcal{K}_{\boldsymbol{u}} \hookrightarrow \mathcal{K}_{\boldsymbol{v}}$ induces a map $\iota_*^{\boldsymbol{u},\boldsymbol{v}} : H_q(\mathcal{K}_{\boldsymbol{u}}, \mathbb{F}) \to H_q(\mathcal{K}_{\boldsymbol{v}}, \mathbb{F})$ from the qth homology of $\mathcal{K}_{\boldsymbol{u}}$ to the qth homology of $\mathcal{K}_{\boldsymbol{v}}$ over a field $\mathbb{F}$. Given a multifiltration $\{\mathcal{K}_{\boldsymbol{u}}\}_{\boldsymbol{u}\in\mathbb{Z}^n}$, the *multiparameter persistence module* is the graded $\mathbb{F}[x_1, \ldots, x_n]$-module $\bigoplus_{\boldsymbol{u}\in\mathbb{Z}^n} H_q(\mathcal{K}_{\boldsymbol{u}}, \mathbb{F})$. The action of x_i on a homogeneous element $\gamma \in H_q(\mathcal{K}_{\boldsymbol{u}}, \mathbb{F})$ is given by $x_i\gamma = \iota^*_{\boldsymbol{u},\boldsymbol{v}}(\gamma)$, where $v_j = u_j + \delta_{ij}$.

Remark 3.2.10. Some researchers define multifiltrations more generally as functors $\mathcal{F} : \mathcal{P} \to \mathit{Simp}$, where $\mathcal{P}$ is any poset and Simp is the category of simplicial complexes, with simplicial maps as morphisms. Definition 3.2.9 is the specific case in which $\mathcal{P} = \mathbb{Z}^n$ and $\mathcal{F}_{\boldsymbol{u}\leq\boldsymbol{v}} : \mathcal{K}_{\boldsymbol{u}} \to \mathcal{K}_{\boldsymbol{v}}$ is an inclusion map.

To see why a fibered filtration function does not typically induce a multifiltration, consider a fibered filtration function $\{f_p : \mathcal{K}^p \to \mathbb{R}\}_{p\in B}$ with $B = \mathbb{R}^n$. Let $\mathcal{K}^p_r := \{\sigma \in \mathcal{K}^p \mid f_p(\sigma) \leq r\}$ denote the r-sublevel set of f_p. It is not necessarily the case that $\mathcal{K}^{p_1}_s \subseteq \mathcal{K}^{p_2}_r$ whenever

$r \leq s$ and $p_1 \leq p_2$. Moreover, there are no canonical simplicial maps $\mathcal{K}_s^{p_1} \to \mathcal{K}_r^{p_2}$, so it is not guaranteed that $\{\mathcal{K}_r^p\}_{(p,r)\in\mathbb{R}^n\times\mathbb{R}}$ is a multifiltration even in the general sense of Remark 3.2.10. Therefore, such a set of filtered complexes cannot be analyzed using multiparameter persistent homology.

3.2.2.2 Fibered barcodes

Consider a multifiltration $\{\mathcal{K}_{\boldsymbol{u}}\}_{\boldsymbol{u}\in\mathbb{R}^n}$. Let $\overline{\mathcal{L}}$ denote the space of lines in $\mathbb{R}^n$ with a parameterization of the form

$$\begin{aligned} &L : \mathbb{R} \to \mathbb{R}^n\,, \\ &L(r) = r\boldsymbol{v} + \boldsymbol{b}\,, \qquad \boldsymbol{v} \in [0,\infty)^n, \|\boldsymbol{v}\| = 1,\, \boldsymbol{b} \in \mathbb{R}^n. \end{aligned}$$

For example, when $n = 2$, the space $\overline{\mathcal{L}}$ is the space of lines in $\mathbb{R}^2$ with non-negative slope, including vertical lines. For each line $L \in \overline{\mathcal{L}}$, we define $\mathcal{K}_r^L := \mathcal{K}_{L(r)}$. That is, $\{\mathcal{K}_r^L\}_{r\in\mathbb{R}}$ is the filtered complex obtained by restricting the multifiltration $\{\mathcal{K}_{\boldsymbol{u}}\}_{\boldsymbol{u}\in\mathbb{R}^n}$ to the line $L \subseteq \mathbb{R}^n$. The set $\{\mathcal{K}_r^L\}_{r\in\mathbb{R}}$ is a filtered complex because $L(r)_i \leq L(s)_i$ for all $r \leq s$ and $i \in \{1,\ldots,n\}$. The *fibered barcode* [And13] is the map that sends $L \in \overline{\mathcal{L}}$ to the barcode for the persistent homology of $\{\mathcal{K}_r^L\}_{r\in\mathbb{R}}$.

A fibered barcode is a PD bundle whose base space is $B = \overline{\mathcal{L}}$. For $L \in \overline{\mathcal{L}}$, the filtration function is

$$\begin{aligned} &f_L : \mathcal{K}^L \to \mathbb{R}\,, \\ &f_L(\sigma) = \min\{r \mid \sigma \in \mathcal{K}_{L(r)}\}\,, \end{aligned}$$

where $\mathcal{K}^L := \bigcup_{r\in\mathbb{R}} \mathcal{K}_{L(r)}$. Unlike the other examples in Section 3.2.1, the simplicial complex $\mathcal{K}^L$ is not independent of $L \in \overline{\mathcal{L}}$.

3.3 A Stratification of the Base Space

There are many different notions of a stratified space [Wei94]. In the present thesis, what we mean by a stratification is the following definition.

Definition 3.3.1. A *stratification* of a topological space B is a nested sequence

$$B^0 \subseteq B^1 \subseteq B^2 \subseteq \cdots \subseteq B^n := B$$

of closed subsets B^m such that the following hold:

1. For all m, the space $B^m \setminus B^{m-1}$ is either empty or a smooth m-dimensional submanifold of B (where we set $B^{-1} := \emptyset$). The *m-dimensional strata* are the connected components of $B^m \setminus B^{m-1}$. We denote the set of strata by $\mathcal{Y}$.

2. The set $\mathcal{Y}$ of strata is *locally finite*: every $p \in B$ has an open neighborhood U such that U intersects finitely many elements of $\mathcal{Y}$.

3. The set $\mathcal{Y}$ of strata satisfy the *Axiom of the Frontier*: If Y_α, $Y_\beta \in \mathcal{Y}$ are strata such that $Y_\beta \cap \overline{Y_\alpha}$, then $Y_\beta \subseteq \overline{Y_\alpha}$. We write that Y_β is a *face* of Y_α.

In the present thesis, B is the base space of a fibered filtration function.

Theorem 3.3.15 says that for any "generic" smooth fibered filtration function (see Section 3.3.2), the base space B can be stratified so that in each stratum $Y \subseteq B$, the set of (birth, death) simplex pairs is constant and can be used to obtain $\mathrm{PD}(f_p)$ for any $p \in Y$.

3.3.1 Piecewise-linear fibered filtrations

As a warm-up, we first consider piecewise-linear fibered filtration functions, which will provide intuition for the general case. However, note that Proposition 3.3.5 below is not simply a special case of Theorem 3.3.15. We consider *all* piecewise-linear fibered filtrations, rather than only generic piecewise-linear fibered filtrations.

First, we establish some notation and definitions.

Definition 3.3.2. An *open half-space* of an affine space A is one of the two connected components of $A \setminus H$ for some hyperplane H.

For example, an open half-space of $\mathbb{R}^n$ is a set of the form $\{\boldsymbol{x} \in \mathbb{R}^n \mid A\boldsymbol{x} > \boldsymbol{b}\}$ for some $n \times n$ matrix A and some vector $\boldsymbol{b} \in \mathbb{R}^n$.

Definition 3.3.3. An *open polyhedron* is the intersection of open half-spaces.

For example, an open polygon P (a polygon without its faces) in $\mathbb{R}^2$ is an open 2D polyhedron because it is the intersection of half-spaces of $\mathbb{R}^2$. The 1D faces (i.e., edges) of P are 1D polyhedra because an edge is a subset of a line $L \subseteq \mathbb{R}^2$ and the edge is the intersection of two half-spaces of L. The 0D faces (i.e., vertices) of P are 0D polyhedra.

We fix a simplicial complex $\mathcal{K}$ for the remainder of this section. For each pair (σ, τ) of simplices in $\mathcal{K}$, we define

$$I(\sigma, \tau) := \{p \in B \mid f(\sigma, p) = f(\tau, p)\}. \tag{3.2}$$

Lemma 3.3.4. Suppose that $f : \mathcal{K} \times B \to \mathbb{R}$ is a continuous fibered filtration function (i.e., $f(\sigma, \cdot) : B \to \mathbb{R}$ is continuous for all simplices $\sigma \in \mathcal{K}$) and that Y is a path-connected subset of B. If all pairs (σ, τ) of simplices satisfy either $I(\sigma, \tau) \cap Y = \emptyset$ or $I(\sigma, \tau) \cap Y = Y$, then the simplex ordering is constant in Y. That is, there is a strict partial order $\prec_Y$ on $\mathcal{K}$ such that for all $y \in Y$, we have that $\prec_{f(\cdot,y)}$ is the same as $\prec_Y$.

Proof. Let (σ, τ) be a pair of simplices.

Case 1: If $I(\sigma, \tau) \cap Y = Y$, then $f(\sigma, p) = f(\tau, p)$ for all $p \in Y$, so $\sigma \not\prec_{f(\cdot,p)} \tau$ and $\tau \not\prec_{f(\cdot,p)} \sigma$ for all $p \in Y$.

Case 2: If $I(\sigma, \tau) \cap Y = \emptyset$, then $f(\sigma, p) \neq f(\tau, p)$ for all $p \in Y$. Let p_0 be a point in Y. Without loss of generality, $\tau \prec_{f(\cdot,p_0)} \sigma$. Therefore, $f(\sigma, p_0) > f(\tau, p_0)$. To obtain a

contradiction, suppose that $f(\sigma, p_1) < f(\tau, p_1)$ for some $p_1 \in Y$. Let $\gamma : [0,1] \to Y$ be a continuous path from p_0 to p_1, and let $g(s) = f(\sigma, \gamma(s)) - f(\tau, \gamma(s))$ for $s \in [0,1]$. By the Intermediate Value Theorem, there is an $s_* \in [0,1]$ such that $f(\sigma, \gamma(s_*)) = f(\tau, \gamma(s_*))$, but this is a contradiction. Therefore, $f(\sigma, p) > f(\tau, p)$ for all $p \in Y$, which implies that $\tau \prec_{f(\cdot,p)} \sigma$ for all $p \in Y$. □

Proposition 3.3.5. If $f : \mathcal{K} \times B \to \mathbb{R}$ is a piecewise-linear fibered filtration function, then B can be partitioned into disjoint polyhedra P on which the simplex ordering induced by f is constant. That is, there is a strict partial order $\prec_P$ on $\mathcal{K}$ such that $\prec_{f(\cdot,p)}$ is the same as $\prec_P$ for all $p \in P$. Consequently, the set $\{(\sigma_b, \sigma_d)\}$ of (birth, death) simplex pairs for f is constant in each P and for any $p \in P$, the persistence diagram $\mathrm{PD}(f_p)$ consists of the diagonal and the multiset $\{(f(\sigma_b), f(\sigma_d)\}$.

Proof. Let Δ be an n-dimensional simplex of the simplicial complex B and let σ and τ be distinct simplices of $\mathcal{K}$. Because $f(\sigma, \cdot)|_\Delta$ and $f(\tau, \cdot)|_\Delta$ are linear, the set $I(\sigma, \tau) \cap \Delta$ is one of the following:

1. the intersection of an $(n-1)$-dimensional hyperplane with Δ;
2. $\emptyset$;
3. Δ;
4. a vertex of Δ.

Therefore, the set $\partial\Delta \bigcup \{I(\sigma, \tau) \cap \Delta \mid \emptyset \subset (I(\sigma, \tau) \cap \Delta) \subset \Delta\}_{\sigma,\tau \in \mathcal{K}}$ partitions Δ into polyhedra. By Lemma 3.3.4, the simplex ordering induced by f is constant on each polyhedron. The last statement of Proposition 3.3.5 follows from Lemma 2.4.4. □

For example, if B is a triangulated surface, then the set

$$L := \bigcup_{\Delta \in B} \partial\Delta \cup \{I(\sigma, \tau) \cap \Delta \mid \emptyset \subset (I(\sigma, \tau) \cap \Delta) \subset \Delta\}_{\sigma,\tau \in \mathcal{K}} \tag{3.3}$$

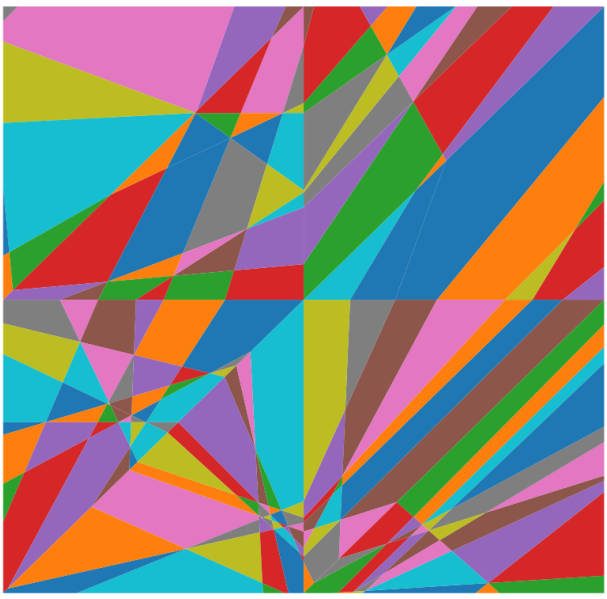


Figure 3.1: A line arrangement that represents the partition of a triangulated surface B (the base space) into polyhedra on which the simplex ordering is constant.

partitions Δ into polyhedra such that the simplex ordering is constant on each polyhedron, including the 1D polyhedra (i.e., edges) and the 0D polyhedra (i.e., vertices). The polygonal subdivision induced by L is called a *line arrangement* $\mathcal{A}(L)$. For an example of such a line arrangement, see Figure 3.1.

3.3.2 Generic smooth fibered filtrations

We now consider generic smooth fibered filtration functions. Throughout Section 3.3.2, we consider a smooth fibered filtration function of the form $f : \mathcal{K} \times B \to \mathbb{R}$ for some n-dimensional smooth compact manifold B and some simplicial complex $\mathcal{K}$. (The fibered filtration f is *smooth* if $f(\sigma, \cdot) : B \to \mathbb{R}$ is smooth for all $\sigma \in \mathcal{K}$.) To make precise the notion of a "generic" fibered filtration function, we consider perturbations of f of a certain form. Because the filtration value of a simplex σ must be at least as large as the filtration value of

any face τ at all $p \in B$, we consider only perturbations $f_{\boldsymbol{a}} : \mathcal{K} \times B \to \mathbb{R}$ of the form

$$f_{\boldsymbol{a}}(\sigma_i, p) := f(\sigma_i, p) + a_i \,,$$

where $\boldsymbol{a}$ is an element of the set

$$A := \{\boldsymbol{a} \in \mathbb{R}^N \mid a_i \leq a_j \text{ for all } i \leq j\} \tag{3.4}$$

and $\sigma_1, \ldots, \sigma_N$ are the simplices of $\mathcal{K}$, indexed such that $i < j$ if σ_i is a proper face of σ_j. By construction, $f_{\boldsymbol{a}}$ is a fibered filtration function for all $\boldsymbol{a} \in A$.

For each simplex σ_k in $\mathcal{K}$, we define the manifold

$$M_k := \{(p, f(\sigma_k, p)) \mid p \in B\} \subseteq B \times \mathbb{R}$$

and for each $\boldsymbol{a} \in A$,we define the manifold

$$M_{\boldsymbol{a},k} := \{(p, f_{\boldsymbol{a}}(\sigma_k, p)) \mid p \in B\} \subseteq B \times \mathbb{R} \,. \tag{3.5}$$

For each pair (σ_i, σ_j) of simplices in $\mathcal{K}$, we define $I(\sigma_i, \sigma_j)$ as in Equation 3.2. The set $I(\sigma_i, \sigma_j)$ is the projection of $M_i \cap M_j \subseteq B \times \mathbb{R}$ to a subset of B. For each $\boldsymbol{a} \in A$, we define the set

$$I_{\boldsymbol{a}}(\sigma_i, \sigma_j) := \{p \in B \mid f_{\boldsymbol{a}}(\sigma_i, p) = f_{\boldsymbol{a}}(\sigma_j, p)\} \,.$$

We also define

$$E^m := \{I(\sigma_{i_1}, \sigma_{j_1}) \cap \cdots \cap I(\sigma_{i_m}, \sigma_{j_m})\} \,,$$

which is the set of all m-way intersections of sets $I(\sigma_i, \sigma_j)$. For all $\boldsymbol{a} \in A$, we define

$$E^m_{\boldsymbol{a}} := \{I_{\boldsymbol{a}}(\sigma_{i_1}, \sigma_{j_1}) \cap \cdots \cap I_{\boldsymbol{a}}(\sigma_{i_m}, \sigma_{j_m})\} \,. \tag{3.6}$$

Lastly, we define

$$E^{m,k}_{\boldsymbol{a}} := \{I_{\boldsymbol{a}}(\sigma_{i_1}, \sigma_{j_1}) \cap \cdots \cap I_{\boldsymbol{a}}(\sigma_{i_m}, \sigma_{j_m}) \mid i_r, j_r \leq k \text{ for all } r\} \,, \tag{3.7}$$

which is the set of m-way intersections that only involve the simplices $\sigma_1, \ldots, \sigma_k$.

Remark 3.3.6. There are several facts to keep in mind. First, it is not guaranteed that $I_{\boldsymbol{a}}(\sigma_i, \sigma_j)$ is homeomorphic to $I(\sigma_i, \sigma_j)$ even for arbitrarily small $\boldsymbol{a}$. Additionally, the sets $I_{\boldsymbol{a}}(\sigma_i, \sigma_j)$ are not "independent" of each other; a perturbation of $f(\sigma, \cdot)$ for a single simplex σ causes a perturbation of $I(\sigma, \tau)$ for all $\tau \in \mathcal{K}$. Furthermore, not every element of E^m is an $(n-m)$-dimensional submanifold, even generically. For example, if $I(\sigma_{i_2}, \sigma_{i_1})$ and $I(\sigma_{i_3}, \sigma_{i_2})$ are $(n-1)$-dimensional submanifolds that intersect transversely, then $I(\sigma_{i_3}, \sigma_{i_1}) \cap I(\sigma_{i_3}, \sigma_{i_2}) \cap I(\sigma_{i_2}, \sigma_{i_1}) = I(\sigma_{i_3}, \sigma_{i_2}) \cap I(\sigma_{i_2}, \sigma_{i_1})$ is an $(n-2)$-dimensional submanifold, rather than an $(n-3)$-dimensional submanifold. Finally, $\bigcap_{r=1}^m I(\sigma_{i_r}, \sigma_{j_r})$ is not necessarily equal to the projection of $\bigcap_{r=1}^m (M_{i_r} \cap M_{j_r})$ to B. In other words, not every intersection in B lifts to an intersection of the manifolds $\{M_k\}_{k=1}^N \subseteq B \times \mathbb{R}$. These are the main subtleties in the proof of Theorem 3.3.15.

Definition 3.3.7. Let S be an element of $E^m_{\boldsymbol{a}}$, where $E^m_{\boldsymbol{a}}$ is defined as in Eq. 3.6. The set S is *m-reduced* if it equals a set of the form $I_{\boldsymbol{a}}(\sigma_{i_1}, \sigma_{j_1}) \cap \cdots \cap I_{\boldsymbol{a}}(\sigma_{i_m}, \sigma_{j_m})$, where $i_1 > i_2 > \cdots > i_m$ and $i_r > j_r$ for all r.

For example, if σ_{i_1}, σ_{i_2}, and σ_{i_3} are distinct simplices, then $I_{\boldsymbol{a}}(\sigma_{i_3}, \sigma_{i_2}) \cap I_{\boldsymbol{a}}(\sigma_{i_2}, \sigma_{i_1})$ is 2-reduced, but $I_{\boldsymbol{a}}(\sigma_{i_3}, \sigma_{i_1}) \cap I_{\boldsymbol{a}}(\sigma_{i_3}, \sigma_{i_2}) \cap I_{\boldsymbol{a}}(\sigma_{i_2}, \sigma_{i_1})$ is not 3-reduced. We define

$$\overline{E^m_{\boldsymbol{a}}} := \{S \in E^m_{\boldsymbol{a}} \mid S \text{ is } m\text{-reduced}\}, \tag{3.8}$$

$$\overline{E^{m,k}_{\boldsymbol{a}}} := \{S \in E^{m,k}_{\boldsymbol{a}} \mid S \text{ is } m\text{-reduced}\}, \tag{3.9}$$

where $E^m_{\boldsymbol{a}}$ is defined as in Equation 3.6 and $E^{m,k}_{\boldsymbol{a}}$ is defined as in Equation 3.7.

Lemma 3.3.8. For all $m \geq 1$, all k, and all $\boldsymbol{a} \in A$, where A is defined as in Eq. 7.12, every element S in $E^{m,k}_{\boldsymbol{a}}$ belongs to $\overline{E^{m',k}_{\boldsymbol{a}}}$ for some $m' \leq m$, where $E^{m,k}_{\boldsymbol{a}}$ and $\overline{E^{m',k}_{\boldsymbol{a}}}$ are defined as in Eqs. 3.7 and 3.9, respectively.

Proof. We prove the lemma by induction on m. For all k, every $S \in E^{1,k}_{\boldsymbol{a}}$ is 1-reduced by definition. Assume that Lemma 3.3.8 is true for $m-1 \geq 1$, and let S be an element of $E^{m,k}_{\boldsymbol{a}}$.

The set S is equal to a set of the form

$$I_{\boldsymbol{a}}(\sigma_{i_1}, \sigma_{j_1}) \cap \cdots \cap I_{\boldsymbol{a}}(\sigma_{i_m}, \sigma_{j_m}) \,,$$

where $i_r > j_r$ for all r and $k \geq i_1 \geq i_2 \geq \cdots \geq i_m$ without loss of generality. By the induction hypothesis,

$$I_{\boldsymbol{a}}(\sigma_{i_2}, \sigma_{j_2}) \cap \cdots \cap I_{\boldsymbol{a}}(\sigma_{i_m}, \sigma_{j_m}) = I_{\boldsymbol{a}}(\sigma_{i'_2}, \sigma_{j'_2}) \cap \cdots \cap I_{\boldsymbol{a}}(\sigma_{i'_\ell}, \sigma_{j'_\ell})$$

for some $\ell \leq m$, where $i'_r > j'_r$ for all r and $k \geq i_1 \geq i_2 \geq i'_2 > i'_3 > \cdots > i'_\ell$. If $i_1 > i'_2$, then S is an element of $E^{\ell,k}$ and we are done. Otherwise,

$$I_{\boldsymbol{a}}(\sigma_{i_1}, \sigma_{j_1}) \cap I_{\boldsymbol{a}}(\sigma_{i'_2}, \sigma_{j'_2}) = I_{\boldsymbol{a}}(\sigma_{i_1}, \sigma_{j_1}) \cap I_{\boldsymbol{a}}(\sigma_{j_1}, \sigma_{j'_2})$$

because $i_1 = i'_2$. If $j_1 = j'_2$, then

$$S = I_{\boldsymbol{a}}(\sigma_{i_1}, \sigma_{j_1}) \cap I_{\boldsymbol{a}}(\sigma_{i'_3}, \sigma_{j'_3}) \cap \cdots \cap I_{\boldsymbol{a}}(\sigma_{i'_\ell}, \sigma_{j'_\ell}) \,,$$

so S is $(\ell - 1)$-reduced. Otherwise,

$$S = I_{\boldsymbol{a}}(\sigma_{i_1}, \sigma_{j_1}) \cap I_{\boldsymbol{a}}(\sigma_{j_1}, \sigma_{j'_2}) \cap I_{\boldsymbol{a}}(\sigma_{i'_3}, \sigma_{j'_3}) \cap \cdots \cap I_{\boldsymbol{a}}(\sigma_{i'_\ell}, \sigma_{j'_\ell}) \,,$$

where $k \geq i_1 > j_1, j'_2$ and $i_1 > i'_r$ for all $r \geq 3$. By the induction hypothesis, the set $I_{\boldsymbol{a}}(\sigma_{j_1}, \sigma_{j'_2}) \cap I_{\boldsymbol{a}}(\sigma_{i'_3}, \sigma_{j'_3}) \cap \cdots \cap I_{\boldsymbol{a}}(\sigma_{i'_\ell}, \sigma_{j'_\ell})$ belongs to $\overline{E_{\boldsymbol{a}}^{\ell',k-1}}$ for some $\ell' \leq \ell - 1$, so S belongs to $\overline{E_{\boldsymbol{a}}^{\ell'+1,k}}$, where $\ell' + 1 \leq \ell \leq m$. □

Lemma 3.3.9. For almost every $\boldsymbol{a} \in A$ (where A is defined as in Eq. 7.12), we have that $M_{\boldsymbol{a},i}$ intersects $M_{\boldsymbol{a},j}$ transversely[2] for $i \neq j$ and every $S \in \overline{E_{\boldsymbol{a}}^m}$ is either $\emptyset$ or an $(n-m)$-dimensional submanifold of B for all $m \in \{1, \ldots, n\}$, where $M_{\boldsymbol{a},i}$ is defined as in Eq. 3.5 and $\overline{E_{\boldsymbol{a}}^m}$ is defined as in Eq. 3.8.

Proof. Define $g^{ij}(p) := f(\sigma_i, p) - f(\sigma_j, p)$ for all $i \neq j$. For almost every $\boldsymbol{a} \in A$, the quantity $a_j - a_i$ is a regular value of g^{ij} by Sard's Theorem. The set of regular values is open for all

[2]When manifolds M_1 and M_2 intersect transversely, we will use the notation $M_1 \pitchfork M_2$.

$i \neq j$ because g^{ij} is smooth and B is compact. Therefore, there is an ϵ^* such that for all $i \neq j$, every $y \in (a_j - a_i - 2\epsilon^*, a_j - a_i + 2\epsilon^*)$ is a regular value of g^{ij}.

Given an $\boldsymbol{a}$ and ϵ^* as above, it suffices to show that for almost every $\boldsymbol{\epsilon} \in \mathbb{R}^N$ with $|\epsilon_i| \leq \epsilon^*$, we have that every $S \in \overline{E^m_{\boldsymbol{a}+\boldsymbol{\epsilon}}}$ is an $(n-m)$-dimensional submanifold of B for all m. For $m = 1$, every element of $\overline{E^1_{\boldsymbol{a}+\boldsymbol{\epsilon}}}$ is of the form $I_{\boldsymbol{a}+\boldsymbol{\epsilon}}(\sigma_i, \sigma_j)$ for some $i \neq j$. The set $I_{\boldsymbol{a}+\boldsymbol{\epsilon}}(\sigma_i, \sigma_j)$ is the $(a_j - a_i + \epsilon_j - \epsilon_i)$-level set of g^{ij}. Because $(a_j - a_i + \epsilon_j - \epsilon_i)$ is a regular value of g^{ij}, the set $I_{\boldsymbol{a}+\boldsymbol{\epsilon}}(\sigma_i, \sigma_j)$ is an $(n-1)$-dimensional submanifold of B and the manifolds $M_{\boldsymbol{a},i}$ and $M_{\boldsymbol{a},j}$ must intersect transversely.

For $m \geq 2$, observe that

$$\overline{E^m_{\boldsymbol{a}}} = \overline{E^{m,2}_{\boldsymbol{a}}} \cup \Big(\bigcup_{k=3}^{N} \overline{E^{m,k}_{\boldsymbol{a}}} \setminus \overline{E^{m,k-1}_{\boldsymbol{a}}} \Big) ,$$

where $\overline{E^{m,k}_{\boldsymbol{a}}}$ is defined as in Eq. 3.9. We induct on $k \in \{2, \ldots, N\}$, where N is the number of simplices in $\mathcal{K}$. When $k = 2$, we have

$$\overline{E^{m,2}_{\boldsymbol{a}+\boldsymbol{\epsilon}}} = \begin{cases} \{I_{\boldsymbol{a}+\boldsymbol{\epsilon}}(\sigma_1, \sigma_2)\} , & m = 1 \\ \emptyset , & m \geq 2 , \end{cases}$$

so every $S \in \overline{E^{m,2}_{\boldsymbol{a}+\boldsymbol{\epsilon}}}$ is either $\emptyset$ or an $(n-m)$-dimensional submanifold of B. Now suppose that $k > 2$ and that every element of $S \in \overline{E^{m,k-1}_{\boldsymbol{a}+\boldsymbol{\epsilon}}}$ is either $\emptyset$ or an $(n-m)$-dimensional submanifold for all m. Every element S in $\overline{E^{m,k}_{\boldsymbol{a}+\boldsymbol{\epsilon}}} \setminus \overline{E^{m,k-1}_{\boldsymbol{a}+\boldsymbol{\epsilon}}}$ is equal to a set of the form

$$S = I_{\boldsymbol{a}+\boldsymbol{\epsilon}}(\sigma_k, \sigma_\ell) \cap S' ,$$

where $\ell \leq k-1$ and $S' \in \overline{E^{m-1,k-1}_{\boldsymbol{a}+\boldsymbol{\epsilon}}}$. We define the vectors $\boldsymbol{\epsilon}^j := (0, \ldots, 0, \epsilon_j, 0, \ldots, 0)$ and $\boldsymbol{b}^j := \boldsymbol{a} + \boldsymbol{\epsilon} - \boldsymbol{\epsilon}^j$. Note that $I_{\boldsymbol{a}+\boldsymbol{\epsilon}}(\sigma_k, \sigma_\ell) = I_{\boldsymbol{b}+\boldsymbol{\epsilon}^k}(\sigma_k, \sigma_\ell)$ because $b^k_i + \epsilon^k_i = a_i + \epsilon_i$ for all i, and $\overline{E^{m-1,k-1}_{\boldsymbol{a}+\boldsymbol{\epsilon}}} = \overline{E^{m-1,k-1}_{\boldsymbol{b}^k}}$ because $a_i + \epsilon_i = b_i$ for all $i \leq k-1$. Therefore, every $S \in \overline{E^{m,k}_{\boldsymbol{a}+\boldsymbol{\epsilon}}} \setminus \overline{E^{m,k-1}_{\boldsymbol{a}+\boldsymbol{\epsilon}}}$ is equal to a set of the form

$$S = I_{\boldsymbol{b}^k+\boldsymbol{\epsilon}^k}(\sigma_k, \sigma_\ell) \cap S'$$

for some $S' \in \overline{E^{m-1,k-1}_{\boldsymbol{b}^k}}$ and $\ell \leq k-1$. Because $g^{\ell k}$ has no critical values between $b_\ell - b_k - \epsilon_k$ and $b_\ell - b_k$, we have that $I_{\boldsymbol{b}^k+\boldsymbol{\epsilon}^k}(\sigma_k, \sigma_\ell)$ is diffeomorphic to $I_{\boldsymbol{b}^k}(\sigma_k, \sigma_\ell)$ for all $\epsilon_k \in (-\epsilon^*, \epsilon^*)$. In other words, $I_{\boldsymbol{b}^k+\boldsymbol{\epsilon}^k}(\sigma_k, \sigma_\ell)$ is a perturbation of $I_{\boldsymbol{b}^k}(\sigma_k, \sigma_\ell)$ for all $\epsilon_k \in (-\epsilon^*, \epsilon^*)$. By Thom's Transversality Theorem, $I_{\boldsymbol{b}^k+\boldsymbol{\epsilon}^k}(\sigma_k, \sigma_\ell)$ intersects every $S' \in E^{m-1,k-1}_{\boldsymbol{b}^k}$ transversely for almost every $\epsilon_k \in (-\epsilon^*, \epsilon^*)$. This shows that S is either $\emptyset$ or an $(n-m)$-dimensional submanifold of B for almost every $\epsilon_k \in (\epsilon^*, \epsilon^*)$. Because there are finitely many elements in $\overline{E^{m,k}_{\boldsymbol{a}}}$, we must have that every $S \in \overline{E^{m,k}_{\boldsymbol{a}}}$ is either $\emptyset$ or an $(n-m)$-dimensional submanifold of B for almost every $\epsilon_k \in (\epsilon^*, \epsilon^*)$. Induction on k concludes the proof. □

Lemma 3.3.10. For all $\boldsymbol{a} \in A$, with A defined as in Eq. 7.12, define

$$B^n_{\boldsymbol{a}} := B \qquad B^m_{\boldsymbol{a}} := \bigcup_{\ell \leq m} \bigcup_{S \in \overline{E^{n-\ell}_{\boldsymbol{a}}}} S \qquad \text{for } m < n\,. \tag{3.10}$$

If $\boldsymbol{a} \in A$ is such that every $S \in \overline{E^{n-\ell}_{\boldsymbol{a}}}$ is either $\emptyset$ or an ℓ-dimensional smooth submanifold for every $\ell \in \{1, \ldots, n\}$, where $\overline{E^{n-\ell}_{\boldsymbol{a}}}$ is defined as in Eq. 3.8, then $B^m_{\boldsymbol{a}} \setminus B^{m-1}_{\boldsymbol{a}}$ is the disjoint union of smooth m-dimensional manifolds.

Proof. We have that

$$B^m_{\boldsymbol{a}} \setminus B^{m-1}_{\boldsymbol{a}} = \bigcup_{S \in \overline{E^{n-m}_{\boldsymbol{a}}}} \Big(S \setminus \bigcup_{\substack{S' \in \overline{E^{n-\ell}_{\boldsymbol{a}}} \\ \ell \leq m-1}} S' \Big)\,.$$

If $S' \in \overline{E^{n-\ell}_{\boldsymbol{a}}}$ is a subset of $S \in \overline{E^{n-m}_{\boldsymbol{a}}}$, then S' is a closed subset of S. Therefore, $S \setminus \bigcup_{\substack{S' \in \overline{E^{n-\ell}_{\boldsymbol{a}}} \\ \ell \leq m-1}} S'$ is an open subset of the smooth manifold S, which implies that $S \setminus \bigcup_{\substack{S' \in \overline{E^{n-\ell}_{\boldsymbol{a}}} \\ \ell \leq m-1}} S'$ is a smooth manifold. If S_1 and S_2 are distinct elements of $\overline{E^{n-m}_{\boldsymbol{a}}}$, then

$$\Big(S_1 \setminus \bigcup_{\substack{S' \in \overline{E^{n-\ell}_{\boldsymbol{a}}} \\ \ell \leq m-1}} S' \Big) \cap \Big(S_2 \setminus \bigcup_{\substack{S' \in \overline{E^{n-\ell}_{\boldsymbol{a}}} \\ \ell \leq m-1}} S' \Big) = \emptyset\,,$$

which completes the proof. □

For the remainder of Section 3.3.2, let $\{B_{\boldsymbol{a}}^m\}_{m=0}^n$ be defined as in Equation 3.10, and define

$$\mathcal{Y}_{\boldsymbol{a}} = \bigcup_{m=0}^{n} \mathcal{Y}_{\boldsymbol{a}}^m \,, \tag{3.11}$$

where $\mathcal{Y}_{\boldsymbol{a}}^m$ is the set of connected components of $B_{\boldsymbol{a}}^m \setminus B_{\boldsymbol{a}}^{m-1}$ (with $B_{\boldsymbol{a}}^{-1} := \emptyset$).

Lemma 3.3.11. Let A be defined as in Eq. 7.12. If $\boldsymbol{a} \in A$ is such that each $Y \in \mathcal{Y}_{\boldsymbol{a}}$ is a manifold, then the simplex ordering induced by f is constant in each Y. That is, there is a strict partial order $\prec_Y$ on $\mathcal{K}$ such that $\prec_{f_{\boldsymbol{a}}(\cdot,y)}$ is the same as $\prec_Y$ for all $y \in Y$.

Proof. Let $Y \in \mathcal{Y}_{\boldsymbol{a}}$. The set Y is connected by definition. Because Y is a manifold, it is also path-connected. For each pair (σ, τ) of simplices, we have by construction that $Y \cap I_{\boldsymbol{a}}(\sigma, \tau)$ equals either $\emptyset$ or Y. (In fact, this statement holds for all $\boldsymbol{a} \in A$ and does not require Y to be a manifold.) By Lemma 3.3.4, the simplex ordering is constant in Y. □

Lemma 3.3.12. For almost every $\boldsymbol{a} \in A$ (where A is defined as in Eq. 7.12), we have that $\bigcap_{r=1}^m I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r})$ is a submanifold of B and

$$T_p\Big(\bigcap_{r=1}^{m} I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r})\Big) = \bigcap_{r=1}^{m} T_p\Big(I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r})\Big) \tag{3.12}$$

for all points $p \in \bigcap_{r=1}^m I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r})$ and all sets $\{(i_r, j_r)\}_{r=1}^m$ of index pairs such that $\{i_r, j_r\} \neq \{i_s, j_s\}$ if $r \neq s$.

Proof. Because there are finitely many sets of index pairs, it suffices to fix a set $\{(i_r, j_r)\}_{r=1}^m$ of index pairs and show that Eq. 3.12 holds for all $y \in \bigcap_{r=1}^m I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r})$ for almost every $\boldsymbol{a} \in A$. By Lemmas 3.3.8 and 3.3.9, the set $\bigcap_{r=1}^m I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r})$ is a manifold for almost every $\boldsymbol{a} \in A$. By Lemma 3.B.2, there is a finite open cover $\{U_k\}_{k=1}^K$ of B such that for each k, there is a disjoint partition $\bigcup_\ell J_{\ell,k} = \{1, \ldots, m\}$ such that $\{i_r, j_r \mid r \in J_{\ell_1,k}\} \cap \{i_r, j_r \mid r \in J_{\ell_2,k}\} = \emptyset$ if $\ell_1 \neq \ell_2$ and

$$\pi\Big(\bigcap_{r \in J_{\ell,k}} (M_{\boldsymbol{a},i_r} \cap M_{\boldsymbol{a},j_r})\Big) \cap U_k = \bigcap_{r \in J_{\ell,k}} I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r}) \cap U_k$$

for all ℓ, where π is the projection $\pi : B \times \mathbb{R} \to B$. Because the number K of open sets is finite, it suffices to fix U_k and show that Eq. 3.12 holds for all $y \in \bigcap_{r=1} I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r}) \cap U_k$ for almost every $\boldsymbol{a} \in A$.

By Lemma 3.B.3, we have

$$T_p\Big(\bigcap_{r=1}^{m} I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r})\Big) = T_p\Big(\bigcap_{\ell}\bigcap_{r\in J_{\ell,k}} I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r})\Big) = \bigcap_{\ell} T_p\Big(\bigcap_{r\in J_{\ell,k}} \Big(I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r})\Big)\Big)$$

for all $p \in \bigcap_{r=1}^{m} I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r}) \cap U_k$ for almost every $\boldsymbol{a} \in A$. By Lemma 3.B.4, we have

$$\bigcap_{\ell} T_p\Big(\bigcap_{r\in J_{\ell,k}} \Big(I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r})\Big)\Big) = \bigcap_{\ell}\bigcap_{r\in J_{\ell,k}} T_p(I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r})$$

for all $y \in \bigcap_{r=1}^{m} I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r}) \cap U_k$ for almost every $\boldsymbol{a} \in A$. □

For any strict partial order $\prec$ on $\mathcal{K}$, we define

$$\begin{aligned} Z_{\boldsymbol{a}}^{\prec} :=& \{p \in B \mid f_{\boldsymbol{a}}(\sigma, p) < f_{\boldsymbol{a}}(\tau, p) \text{ if } \sigma \prec \tau \\ & \text{and } f_{\boldsymbol{a}}(\sigma, p) = f_{\boldsymbol{a}}(\tau, p) \text{ if } \sigma \not\prec \tau \text{ and } \tau \not\prec \sigma\}\,. \end{aligned} \tag{3.13}$$

That is, $Z_{\boldsymbol{a}}^{\prec}$ is the subset of B such that for all z in $Z_{\boldsymbol{a}}^{\prec}$, the strict partial order $\prec_{f_{\boldsymbol{a}}(\cdot,z)}$ is the same as $\prec$.

Lemma 3.3.13. Let A be defined as in Eq. 7.12. If $\boldsymbol{a} \in A$ is such that

1. every $Y \in \mathcal{Y}_{\boldsymbol{a}}$ is a manifold, where $\mathcal{Y}_{\boldsymbol{a}}$ is defined as in Eq. 3.11,
2. $M_{\boldsymbol{a},i} \pitchfork M_{\boldsymbol{a},j}$ for all $i \neq j$, where $M_{\boldsymbol{a},i}$ is defined as in Eq. 3.5,
3. $\bigcap_{r=1}^{m} I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r})$ is a manifold for all sets $\{(i_r, j_r\}_{r=1}^{m}$ of index pairs, and
4. $T_p\Big(\bigcap_{r=1}^{m} I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r})\Big) = \bigcap_{r=1}^{m} T_p\Big(I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r})\Big)$ for all sets $\{(i_r, j_r\}_{r=1}^{m}$ of index pairs and all $p \in \bigcap_{r=1}^{m} I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r})$,

then $\mathcal{Y}_{\boldsymbol{a}}$ is locally finite.

Proof. Let p be a point in B. There are finitely many strict partial orders $\prec_1, \ldots, \prec_i$ on $\mathcal{K}$. By Lemma 3.B.5, we have that for each $j \in \{1, \ldots, i\}$, there is a subset $\mathcal{Y}_{\boldsymbol{a}}^{\prec_j} \subseteq \mathcal{Y}_{\boldsymbol{a}}$ such that $Z_{\boldsymbol{a}}^{\prec_j} = \bigcup_{Y \in \mathcal{Y}_{\boldsymbol{a}}^{\prec_j}} Y$. For each j, the point p has a neighborhood U_j that intersects at most one $Y \in \mathcal{Y}_{\boldsymbol{a}}^{\prec_j}$ by Lemma 3.B.6. Therefore $\bigcap_{j=1}^{i} U_j$ is a neighborhood of p that intersects at most i elements of $\mathcal{Y}_{\boldsymbol{a}}$. □

Lemma 3.3.14. Let A be defined as in Eq. 7.12. If $\boldsymbol{a} \in A$ is such that

1. every $S \in \overline{E_{\boldsymbol{a}}^{n-\ell}}$ is an ℓ-dimensional smooth submanifold for every $\ell \in \{1, \ldots, n\}$, where $\overline{E_{\boldsymbol{a}}^{n-\ell}}$ is defined as in Eq. 3.8,
2. $M_{\boldsymbol{a},i} \pitchfork M_{\boldsymbol{a},j}$ for all $i \neq j$, where $M_{\boldsymbol{a},i}$ is defined as in Eq. 3.5,
3. $\bigcap_{i=r}^{m} I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r})$ is a manifold for all sets $\{(i_r, j_r\}_{r=1}^{m}$ of index pairs, and
4. $T_p\Big(\bigcap_{i=r}^{m} I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r})\Big) = \bigcap_{i=1}^{m} T_p\Big(I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r})\Big)$ for all sets $\{(i_r, j_r\}_{r=1}^{m}$ of index pairs and all $p \in \bigcap_{i=r}^{m} I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r})$,

then $\mathcal{Y}_{\boldsymbol{a}}$ satisfies the axiom of the frontier in Definition 3.3.1, where $\mathcal{Y}_{\boldsymbol{a}}$ is defined as in Eq. 3.11

Proof. By Lemma 3.3.10, each $Y \in \mathcal{Y}_{\boldsymbol{a}}$ is a manifold. Let Y_α be an element of $\mathcal{Y}_{\boldsymbol{a}}$. It suffices to show that if $Y_\beta \neq Y_\alpha$ is another element of $\mathcal{Y}_{\boldsymbol{a}}$ and $Y_\beta \cap \partial Y_\alpha \neq \emptyset$, where ∂Y_α denotes the boundary of the manifold Y_α, then $Y_\beta \subseteq \partial Y_\alpha$.

By Lemma 3.3.11, the simplex ordering induced by f is constant on each Y, so there is a strict partial order $\prec_\alpha$ on $\mathcal{K}$ such that $\prec_{f_{\boldsymbol{a}}(\cdot, y)}$ is the same as $\prec_\alpha$ for all $y \in Y_\alpha$. Let $Z_{\boldsymbol{a}}^{\prec_\alpha}$ be defined as in Equation 3.13. By Lemma 3.B.5, there is a subset $\mathcal{Y}_{\boldsymbol{a}}^{\prec_\alpha} \subseteq \mathcal{Y}_{\boldsymbol{a}}$ such that $Y_\alpha \in \mathcal{Y}_{\boldsymbol{a}}^{\prec_\alpha}$ and $Z_{\boldsymbol{a}}^{\prec_\alpha} = \bigcup_{Y \in \mathcal{Y}_{\boldsymbol{a}}^{\prec_\alpha}} Y$. We have $\partial Z_{\boldsymbol{a}}^{\prec_\alpha} = \bigcup_{Y \in \mathcal{Y}_{\boldsymbol{a}}^{\prec_\alpha}} \partial Y$ because by Lemma 3.3.13, $\mathcal{Y}$ is locally finite. Therefore,

$$Y_\beta \cap \partial Z_{\boldsymbol{a}}^{\prec_\alpha} = \bigcup_{Y \in \mathcal{Y}_{\boldsymbol{a}}^{\prec_\alpha}} (Y_\beta \cap \partial Y)\,. \tag{3.14}$$

By Lemmas 3.B.5 and 3.B.7, we have that if $Y \in \mathcal{Y}_{\boldsymbol{a}}$ intersects $\partial Z_{\boldsymbol{a}}^{\prec\alpha}$, then $Y \subseteq \partial Z_{\boldsymbol{a}}^{\prec\alpha}$. Therefore $Y_\beta \subseteq \partial Z_{\boldsymbol{a}}^{\prec\alpha}$ because $Y_\beta \cap \partial Z_{\boldsymbol{a}}^{\prec\alpha} \supseteq Y_\beta \cap \partial Y_\alpha \neq \emptyset$. Together with Equation 3.14, this shows that

$$Y_\beta = \bigcup_{Y \in \mathcal{Y}_{\boldsymbol{a}}^{\prec\alpha}} (Y_\beta \cap \partial Y)\,. \tag{3.15}$$

By Lemma 3.B.6, every point in B has a neighborhood that intersects at most one $Y \in \mathcal{Y}_{\boldsymbol{a}}^{\prec\alpha}$, so $\partial Y' \cap \partial Y = \emptyset$ for all $Y, Y' \in \mathcal{Y}_{\boldsymbol{a}}^{\prec\alpha}$ such that $Y \neq Y'$. Because Y_β is connected (by definition) and $Y_\beta \cap \partial Y_\alpha \neq \emptyset$, we must have that $Y_\beta \cap \partial Y = \emptyset$ for all $Y \in \mathcal{Y}_{\boldsymbol{a}}^{\prec\alpha}$ such that $Y \neq Y_\alpha$. By Equation 3.15,

$$Y_\beta = Y_\beta \cap \partial Y_\alpha \subseteq \partial Y_\alpha\,.$$

□

Theorem 3.3.15. *Let B be a smooth compact n-dimensional manifold. For every $\boldsymbol{a} \in A$, define $\{B_{\boldsymbol{a}}^m\}_{m=0}^n$ as in Equation 3.10, with A defined as in Eq. 7.12. For almost every $\boldsymbol{a} \in A$, we have that $\{B_{\boldsymbol{a}}^m\}_{m=0}^n$ is a stratification of B. In each stratum Y, the simplex ordering induced by $f_{\boldsymbol{a}}$ is constant. (In other words, there is a strict partial order $\prec_Y$ on $\mathcal{K}$ such that $\prec_{f_{\boldsymbol{a}}(\cdot,y)}$ is the same as $\prec_Y$ for all $y \in Y$.) Consequently, the set $\{(\sigma_b, \sigma_d)\}$ of (birth, death) simplex pairs is constant in each stratum Y and for any $p \in Y$, the persistence diagram $PD(f_p)$ consists of the diagonal (with infinite multiplicity) and the multiset $\{(f(\sigma_b, p), f(\sigma_d, p))\}$.*

Proof. By Lemmas 3.3.8, 3.3.9, 3.3.10, 3.3.12, 3.3.13, and 3.3.14, $\{B_{\boldsymbol{a}}^m\}_{m=0}^n$ is a stratification of B for almost every $\boldsymbol{a} \in A$. By Lemma 3.3.11, the simplex ordering induced by $f_{\boldsymbol{a}}$ is constant in each stratum $Y \in \mathcal{Y}_{\boldsymbol{a}}$ whenever $\{B_{\boldsymbol{a}}^m\}_{m=0}^n$ is a stratification of B. The last statement of Theorem 3.3.15 follows from Lemma 2.4.4. □

3.4 Sections of PD Bundles

Definition 3.4.1 (Local section)**.** Let (E, B, π) be a PD bundle. A *local section* is a continuous map $s : U \to E$, where U is an open set in B and $\pi \circ s(p) = p$ for all $p \in U$.

For example, consider a vineyard, in which B is an interval I in $\mathbb{R}$. Let (t_0, T) be an open interval in I. A local section in the vineyard is a map $s : (t_0, T) \to E$ that parameterizes an open subset of one of the vines (a curve in $\mathbb{R}^3$).

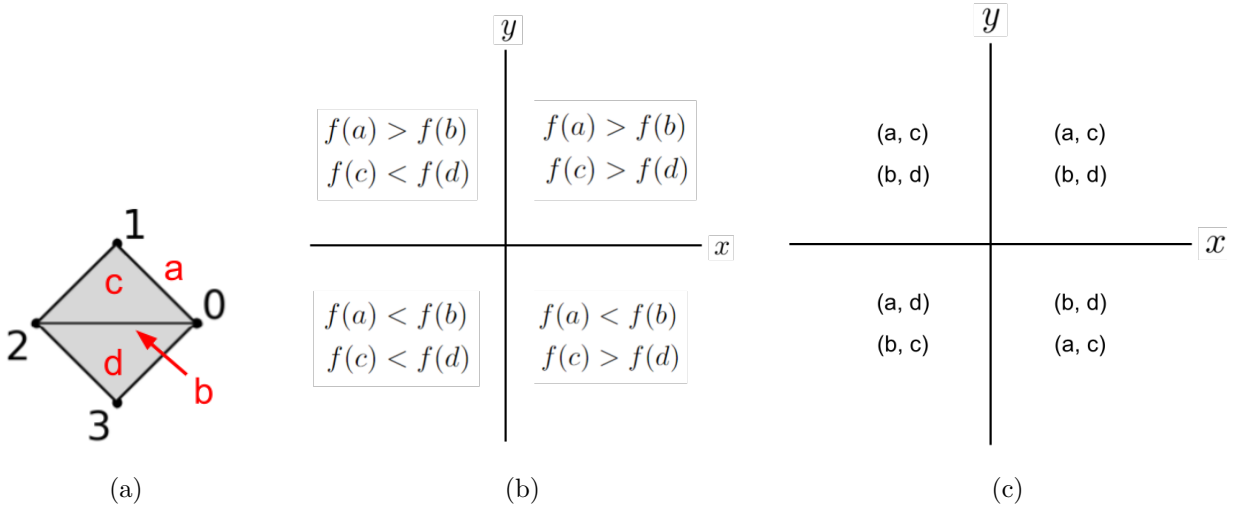
Definition 3.4.2 (Global section)**.** Let (E, B, π) be a PD bundle. A *global section* is a continuous map $s : B \to E$ with $\pi \circ s(p) = p$ for all $p \in B$. In particular, a *nontrivial global section* is a global section $s : B \to E$ such that there exists a $p_* \in B$ for which $s(p_*)$ is not on the diagonal of $PD(f_{p_*})$.

In a vineyard, every local section can be extended to a global section. In other words, we can trace out how the persistence of a single homology class changes over $B = [t_0, t_1] \subseteq \mathbb{R}$, so there are individual "vines" in the vineyard. We will show that local sections of a PD bundle cannot necessarily be extended to global sections. This means that there are not typically individual "vines" in a PD bundle.

Proposition 3.4.3. There is a PD bundle (E, B, π) for which no nontrivial global sections exist.

Proof. The proof is constructive. Let $\mathcal{K}$ be the simplicial complex in Figure 3.2a, which has vertices 0, 1, 2, and 3. Let a be the edge with vertices $(0, 1)$, let b be the edge with vertices $(0, 2)$, let c be the triangle with vertices $(0, 1, 2)$, and let d be the triangle with vertices $(0, 2, 3)$.

Let $f : \mathcal{K} \times \mathbb{R}^2 \to \mathbb{R}$ be a continuous fibered filtration function that satisfies the following



(a) (b) (c)

Figure 3.2: (A) The simplicial complex $\mathcal{K}$ that is defined in the proof of Proposition 3.4.3. (B) The conditions on the fibered filtration $f : \mathcal{K} \times \mathbb{R}^2 \to \mathbb{R}$ that is defined in the proof of Proposition 3.4.3. (C) The (birth, death) simplex pairs in each quadrant for the 1D PH.

conditions:

$$
\begin{aligned}
f(c,(x,y)), f(d,(x,y)) > f(a,(x,y)), f(b,(x,y)) > 0 \qquad & \text{for all } (x,y) \in \mathbb{R}^2\,, \\
f(a,(x,y)) > f(b,(x,y))\,, \qquad & y > 0\,, \\
f(a,(x,y)) < f(b,(x,y))\,, \qquad & y < 0\,, \\
f(c,(x,y)) > f(d,(x,y))\,, \qquad & x > 0\,, \\
f(c,(x,y)) < f(d,(x,y)\,,) \qquad & x < 0\,, \\
f(\sigma,(x,y)) = 0\,, \qquad & \text{for all other } \sigma \text{ for all } x, y\,.
\end{aligned}
$$

The conditions on the fibered filtration function f are illustrated in Figure 3.2b. These conditions imply that simplices a and b swap their order along the x-axis and the simplices c and d swap their order along the y-axis.

In Figure 3.2c, we list the (birth, death) simplex pairs for the 1D PH in each quadrant. In quadrants 1, 2, and 4, the simplex pairs are (a, c) and (b, d). In quadrant 3, the simplex pairs are (a, d) and (b, c).

Let $(E, \mathbb{R}^2, \pi)$ be the corresponding PD bundle, where $E = \{((x,y),z) \mid (x,y) \in \mathbb{R}^2, z \in PD_1(f(\cdot,(x,y)))\}$ is the total space and π is the projection to $\mathbb{R}^2$. We will show that $(E, \mathbb{R}^2, \pi)$ has no nontrivial global sections.

If $s : B \to E$ is a global section and $s(p_*)$ is on the diagonal of $\mathrm{PD}(f(\cdot, p_*))$ for some $p_* \in B$, then s is a trivial section because $f(\sigma_b, p) \neq f(\sigma_d, p)$ for all p for any (birth, death) simplex pairs (σ_b, σ_d) at p. Therefore, if $s : B \to E$ is a nontrivial global section, $s(p)$ is not on the diagonal of $\mathrm{PD}(f(\cdot, p))$ for any $p \in B$.

Suppose that $p : [0,1] \to E$ is a continuous path such that $p(u)$ is not on the diagonal of $\mathrm{PD}(f(\cdot,(x,y)))$ for any $(x,y) \in \mathbb{R}^2$ and such that

$$\pi \circ p(u) = \theta(u) := (\cos(2\pi u + \pi/4), \sin(2\pi u + \pi/4)) \in S^1 \,. \tag{3.16}$$

That is, $\pi \circ p$ is a parameterization of S^1 that starts in the first quadrant of $\mathbb{R}^2$. The path p is determined uniquely by its initial condition $p(0)$. The simplex pairs in the first quadrant are (a,c) and (b,d), so $p(0)$ equals either $(p_0, (f(a,p_0), f(c,p_0)))$ or $(p_0, (f(b,p_0), f(d,p_0)))$, where $p_0 = (\sqrt{2}/2, \sqrt{2}/2)$. In Figure 3.3, we illustrate the two possibilities for the path p. If $p(0) = (p_0, (f(a,p_0), f(c,p_0)))$, then $p(1) = (p_0, (f(b,p_0), f(d,p_0)))$; f $p(0) = (p_0, (f(b,p_0), f(d,p_0)))$, then $p(1) = (p_0, (f(a,p_0), f(c,p_0)))$. In either case, $p(0) \neq p(1)$.

To obtain a contradiction, suppose that there were a nontrivial global section $s : \mathbb{R}^2 \to E$. Let $p : [0,1] \to E$ be the path $p = s \circ \theta$, where θ is the parameterization of S^1 defined in Equation 3.16. Then $p(0) \neq p(1)$ because p is a path satisfying Equation 3.16, but $p(0) = s(\sqrt{2}/2, \sqrt{2}/2) = p(1)$. □

Note that we will use the fibered filtration $f : \mathcal{K} \times \mathbb{R}^2 \to \mathbb{R}$ that was constructed in Proposition 3.4.3 as a running example throughout Section 3.5.2.

Remark 3.4.4. Even when $\dim(B) = 1$, it is not guaranteed that a nontrivial global section exists. To see this, consider the 1D PH of the fibered filtration function above restricted to $S^1 \subseteq \mathbb{R}^2$. In this example, $\dim(B) = 1$ and a nontrivial global section does not exist.

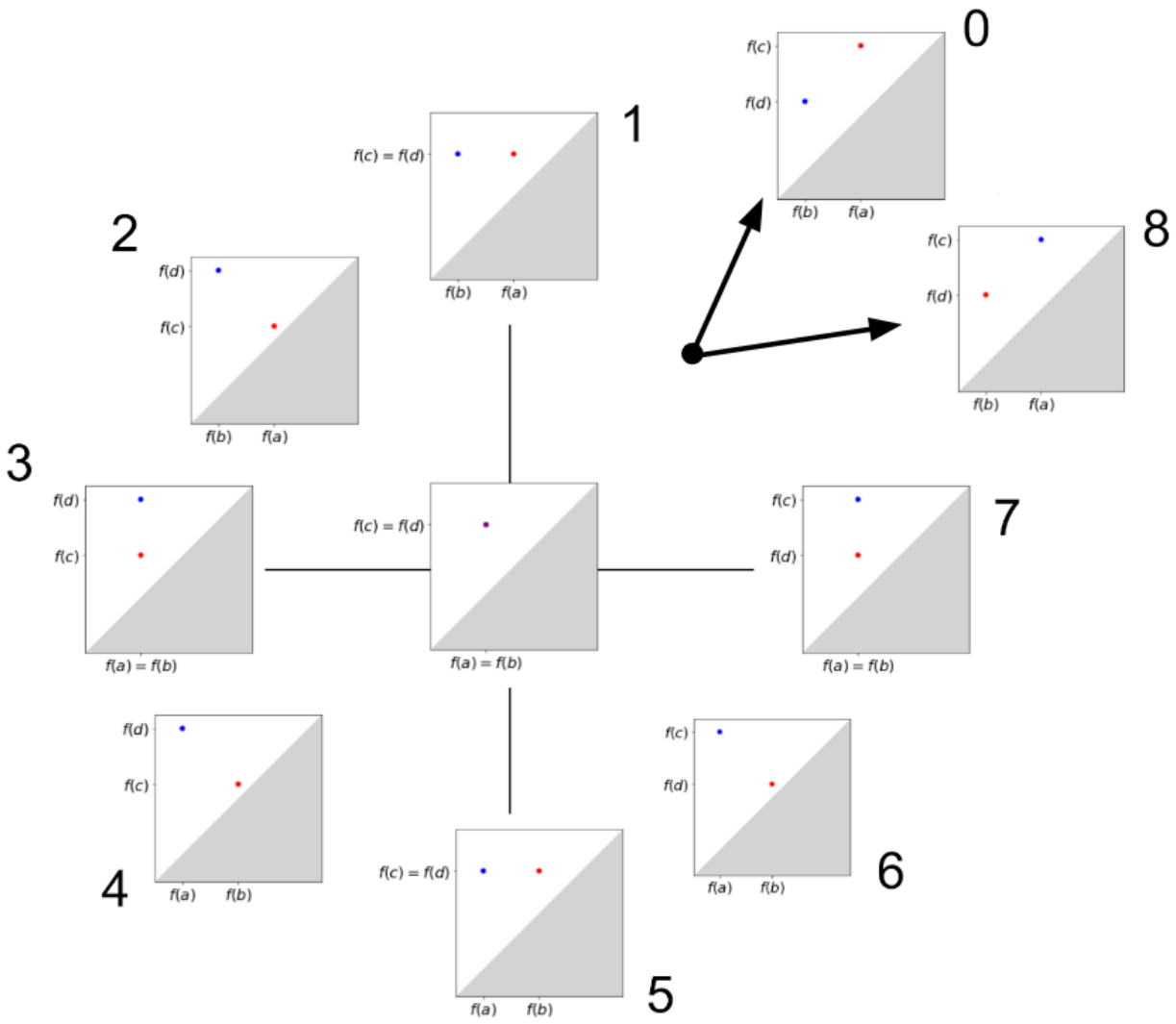


Figure 3.3: A visualization of the two choices for the path $p : [0, 1] \to E$ in the proof of Proposition 3.4.3, where E is the total space of the PD bundle. We show 10 fibers of the PD bundle for various points $p \in \mathbb{R}^2$. The first nine PDs (labeled 0 through 8) are PDs for points $p \in S^1$; the kth PD is the PD at $t_k = \theta(u_k)$, where $u_k = k/8$ and $\theta(u)$ is the parameterization of S^1 given by Equation 3.16. Note that $\theta(0) = (\sqrt{2}/2, \sqrt{2}/2) \in S^1$. The two choices for the path $p(u)$, which depend only on the choice of $p(u_0)$, are shown in red and blue, respectively. For each k, the red (respectively, blue) dot in the kth PD is equal to $p(u_k)$ when $p(u_0)$ is the red (respectively, blue) point in the 0th PD. Observe that $p(u_0) \neq p(u_8)$ even though $p_0 = p_8$. The unlabeled PD at the origin is the PD for the origin in $\mathbb{R}^2$.

Remark 3.4.5. In the example of Proposition 3.4.3, it was the "singularity" (the point $(0,0) \in \mathbb{R}^2$ at which the PD had a point of multiplicity greater than 1) that prevented the existence of a nontrivial global section. Singularities are typical if and only if $\dim(B) > 1$. This is for the following reason. For a generic fibered filtration function of the form $f : \mathcal{K} \times B \to \mathbb{R}$, where B is an n-dimensional compact manifold, Lemma 3.3.9 shows that the intersection $I(\sigma, \tau)$ is an $(n-1)$-dimensional submanifold of B for all simplices $\tau \neq \sigma$ and that $I(\sigma_{i_1}, \sigma_{j_1}) \cap I(\sigma_{i_2}, \sigma_{j_2})$ is an $(n-2)$-dimensional manifold for all distinct $\sigma_{i_1}, \sigma_{i_2}, \sigma_{j_1}, \sigma_{j_2}$. A singularity occurs at $p_* \in B$ when there are two (birth, death) simplex pairs (σ_b^1, σ_d^1), (σ_b^2, σ_d^2) at p_* such that $p_* \in I(\sigma_b^1, \sigma_b^2) \cap I(\sigma_d^1, \sigma_d^2)$. When $\dim(B) \geq 2$, the intersection $I(\sigma_b^1, \sigma_b^2) \cap I(\sigma_d^1, \sigma_d^2)$ is nonempty in the generic case because it is an $(n-2)$-dimensional submanifold of B. When $\dim(B) = 1$, the intersection $I(\sigma_b^1, \sigma_b^2) \cap I(\sigma_d^1, \sigma_d^2)$ is empty in the generic case, so singularities do not typically exist when $\dim(B) = 1$.

3.5 A Compatible Cellular Sheaf

For a given fibered filtration function that induces a stratification of B as in Theorem 3.3.15, we construct a compatible cellular sheaf. We discuss a motivating example in Section 3.5.1, and give the definition in Section 3.5.2.

3.5.1 A motivating example

Again consider the example in the proof of Proposition 3.4.3, and also again consider the path $p : [0,1] \to E$ that is determined uniquely by the choice of

$$p(0) \in \{(p_0, (f(a, p_0), f(c, p_0))), (p_0, (f(b, p_0), f(d, p_0)))\},$$

where $p_0 = (\sqrt{2}/2, \sqrt{2}/2)$. The two possibilities for the path p are illustrated in Figure 3.3. For example, if $p(0) = (p_0, (f(a, p_0), f(c, p_0)))$, then

$$p(u) = \begin{cases} \Big(\theta(u), (f(a, \theta(u)), f(c, \theta(u)))\Big), & u \in [0, 1/8] \\ \Big(\theta(u), (f(a, \theta(u)), f(c, \theta(u)))\Big), & u \in [1/8, 3/8] \\ \Big(\theta(u), (f(b, \theta(u)), f(c, \theta(u)))\Big), & u \in [3/8, 5/8] \\ \Big(\theta(u), (f(b, \theta(u)), f(d, \theta(u)))\Big), & u \in [5/8, 7/8] \\ \Big(\theta(u), (f(b, \theta(u)), f(d, \theta(u)))\Big), & u \in [7/8, 1], \end{cases}$$

where $\theta(u)$ is the parameterization of S^1 given by Equation 3.16. As we move through the quadrants of $\mathbb{R}^2$, the point in the PD that represents the pair (a, c) in the first quadrant becomes the point that represents the pair (a, c) in the second quadrant, which becomes the point that represents the pair (b, c) in the third quadrant, which becomes the point that represents the pair (b, d) in the fourth quadrant, which becomes the point that represents the pair (b, d) in the first quadrant. One can do a similar analysis for the case in which $p(0) = (p_0, (f(b, p_0), f(d, p_0)))$.

This analysis yields a bijection of the (birth, death) simplex pairs for any pair of adjacent quadrants. We illustrate the bijections in Figure 3.4. The bijection between the simplex pairs in a given quadrant and one of its adjacent quadrants is the same as the bijection defined by the update rule of Cohen-Steiner et al. [CEM06] for updating the simplex pairs in a vineyard. A combinatorial perspective on Proposition 3.4.3 is that there is no consistent way of choosing a simplex pair in each quadrant such that if (σ_b, σ_d) is the (birth, death) simplex pair chosen for a given quadrant and (τ_b, τ_d) is the (birth, death) simplex pair chosen for an adjacent quadrant, then (σ_b, σ_d) and (τ_b, τ_d) are matched in the bijection between the two quadrants. This is because if we choose an initial simplex pair in one of the quadrants and then walk in a circle through the other quadrants, then the simplex pair at which we finish is different from the initial simplex pair. For example, if we start at (a, c) in the first quadrant, then we finish at (b, d) when we return to the first quadrant, and vice versa. This

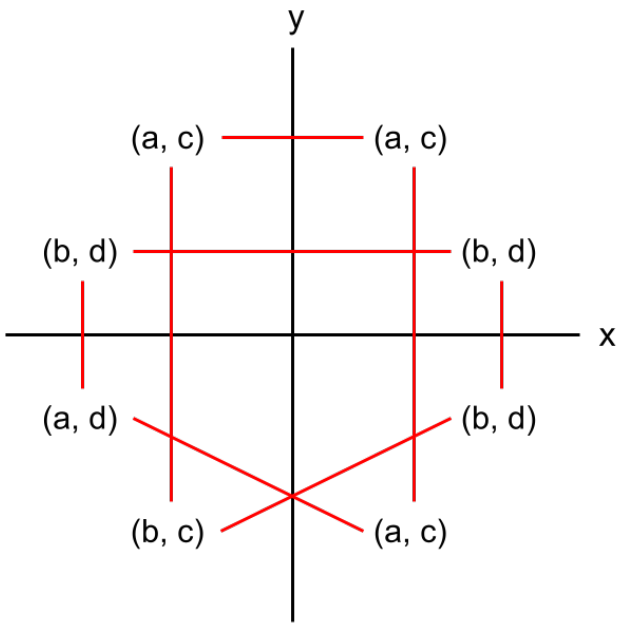


Figure 3.4: The (birth, death) simplex pairs in each quadrant for the 1D PH of the fibered filtration function in the proof of Proposition 3.4.3 (see also Figure 3.2). For each pair of adjacent quadrants, there is a bijection between their sets of simplex pairs; this bijection is equal to the bijection given by the update rule of Cohen-Steiner et al. [CEM06]. The red lines connect simplex pairs that are in bijection with each other.

is a discrete way of illustrating the non-existence of a nontrivial global section.

3.5.2 Definition of a compatible cellular sheaf

I generalize the discussion in Section 3.5.1 to fibered filtration functions of the form $f : \mathcal{K} \times B \to \mathbb{R}$ that have a stratification (see Definition 3.3.1) of B such that in each stratum Y, the simplex ordering that is induced by f is constant. (In other words, there is a strict partial order $\prec_Y$ on $\mathcal{K}$ such that $\prec_{f(\cdot,y)}$ is the same as $\prec_Y$ for all $y \in Y$.) Theorem 3.3.15 guarantees that such a stratification exists for generic fibered filtration functions, and Proposition 3.3.5 guarantees that such a stratification exists for all piecewise-linear fibered filtration functions. We denote the set of strata by $\mathcal{Y} = \{Y_\alpha\}_{\alpha \in J}$ for some index set J.

Definition 3.5.1. Suppose that $\mathcal{F}$ is a **Set**-valued cellular sheaf whose cell complex, stalks, and morphisms are of the following form:

1. **The cell complex:** The cell complex on which $\mathcal{F}$ is constructed is the graph G such that there is a vertex v_α for each stratum $Y_\alpha \in \mathcal{Y}$ and an edge $e_{\beta,\alpha} = (v_\beta, v_\alpha)$ if $Y_\beta \in \mathcal{Y}$ is a face of Y_α. The 0-cells of the cell complex are the vertices of G and the 1-cells are the edges of G.

2. **The stalks:** Let S_α denote the set of (birth, death) simplex pairs for a stratum Y_α. The stalk at a 0-cell $v_\alpha \in G$ is $\mathcal{F}(v_\alpha) := S_\alpha$. For a 1-cell $e_{\beta,\alpha} \in G$, where Y_β is a face of Y_α, the stalk at $e_{\beta,\alpha}$ is $\mathcal{F}(e_{\beta,\alpha}) := S_\alpha$.

3. **The morphisms:** If $Y_\beta \in \mathcal{Y}$ is a face of $Y_\alpha \in \mathcal{Y}$, then the morphism $\mathcal{F}_{v_\beta \leq e_{\beta,\alpha}} : \mathcal{F}(v_\alpha) \to \mathcal{F}(e_{\beta,\alpha})$ is the identity map and the morphism $\mathcal{F}_{v_\beta \leq e_{\beta,\alpha}} : \mathcal{F}(v_\beta) \to \mathcal{F}(e_{\beta,\alpha})$ is

$$\mathcal{F}_{v_\beta \leq e_{\beta,\alpha}} := \phi^{\mathrm{idx}_\beta, \mathrm{idx}_\alpha} ,$$

   where $\phi^{\mathrm{idx}_\beta, \mathrm{idx}_\alpha}$ is of the form in Eq. 2.4 and $\mathrm{idx}_\alpha : \mathcal{K} \to \{1, \ldots, N\}$ and $\mathrm{idx}_\beta : \mathcal{K} \to \{1, \ldots, N\}$ are the simplex indexings (recall Definition 2.4.3) on Y_α and Y_β, respectively.

Then the cellular sheaf $\mathcal{F}$ is a *compatible cellular sheaf* for the fibered filtration function $f : \mathcal{K} \times B \to \mathbb{R}$.

It is not guaranteed that there is a *unique* compatible cellular sheaf for a given fibered filtration function f. Although the cell complex (the graph G) is determined uniquely by f, the stalks and morphisms are not. Recall from Definition 2.4.3 that the simplex indexing that is induced by f may depend on an intrinsic indexing $\sigma_1, \ldots, \sigma_N$ of the simplices in $\mathcal{K}$. (The intrinsic indexing breaks ties when two simplices have the same filtration value.) For a stratum Y_α such that $f(\sigma, y) = f(\tau, y)$ for all $y \in Y_\alpha$ for some pair (σ, τ) of simplices, the simplex indexing $\mathrm{idx}_{f(\cdot, Y_\alpha)}$ depends on the intrinsic indexing, so S_α may not be determined

uniquely by f. If S_α is not determined uniquely by f, then for any face Y_β of Y_α, the stalks $\mathcal{F}(v_\alpha)$ and $\mathcal{F}(e_{\beta,\alpha})$ are not determined uniquely by f. As discussed in Remark 2.6.1, a bijection $\phi^{\text{idx}_\beta,\text{idx}_\alpha}$ of the form in Equation 2.4 is not determined uniquely by f if idx_β and idx_α do not differ by the transposition of two consecutive simplices. Therefore, the morphism $\mathcal{F}_{v_\beta \leq e_{\beta,\alpha}}$ is not necessarily determined uniquely by f.

However, many aspects of the stalks and morphisms *are* determined uniquely by f. Suppose that $Y_\beta \in \mathcal{Y}$ is a face of $Y_\alpha \in \mathcal{Y}$. If $f(\sigma, y) \neq f(\tau, y)$ for all y in Y_α and all simplices $\sigma \neq \tau$, then the simplex indexing $\text{idx}_{f(\cdot,Y_\alpha)}$ is determined uniquely by f, so the stalks $\mathcal{F}(v_\alpha)$ and $\mathcal{F}(e_{\beta,\alpha})$ are determined uniquely by f. Theorem 3.3.15 guarantees that this is the generic case when Y_α is an n-dimensional stratum (where $n = \dim(B)$). There are also conditions under which a morphism is determined uniquely by f. The morphism $\mathcal{F}_{v_\alpha \leq e_{\beta,\alpha}} : S_\alpha \to S_\alpha$ must be the identity map. The morphism $\mathcal{F}_{v_\alpha \leq e_{\beta,\alpha}} := \phi^{\text{idx}_\beta,\text{idx}_\alpha}$ is determined uniquely by f when idx_β and idx_α differ by the transposition of two consecutive simplices. Theorem 3.3.15 guarantees that this is the generic case when Y_β is a "top-dimensional" face of Y_α (i.e., when $\dim(Y_\beta) = \dim(Y_\alpha) - 1$).

Example 3.5.2. Again consider a fibered filtration function $f : \mathcal{K} \times \mathbb{R}^2 \to \mathbb{R}$ of the form defined in Prop. 3.4.3, with $\mathcal{K}$ defined as in Figure 3.2a with $N = 11$ simplices. We construct a compatible cellular sheaf $\mathcal{F}$ as follows.

1. **The cell complex:** The strata are the open quadrants $Q_1, \ldots, Q_4$, the open half-axes $A_{12}, A_{23}, A_{34}, A_{14}$ with $A_{ij} = (\partial Q_i \cap \partial Q_j) \setminus \{\mathbf{0}\}$, and the point $\mathbf{0} \in \mathbb{R}^2$. The associated graph G (the cell complex for $\mathcal{F}$) has a vertex v_{Q_i} for the ith quadrant, a vertex $v_{A_{ij}}$ for the (i,j)th half-axis, and a vertex $v_{\mathbf{0}}$ for the point $\mathbf{0}$. The graph G has edges $(v_{A_{ij}}, v_{Q_i})$ and $(v_{A_{ij}}, v_{Q_j})$ for each half-axis A_{ij}, and it has an edge $(v_{\mathbf{0}}, v)$ for every vertex $v \in G$ such that $v \neq v_{\mathbf{0}}$.

2. **The stalks:** We index the simplices of $\mathcal{K}$ such that $\sigma_8 = a$, $\sigma_9 = b$, $\sigma_{10} = c$, and $\sigma_{11} = d$, where a, b, c, d are the simplices defined in Figure 3.2a. The stalk at v_{Q_1} is

$S_{Q_1} = \{(a,c),(b,d)\}$. The vertices v_{Q_2} and $v_{A_{12}}$ have the same stalk $\{(a,c),(b,d)\}$; the vertices v_{Q_3}, $v_{A_{23}}$, and v_{34} have the same stalk $\{(a,d),(b,c)\}$;and the vertices v_{Q_4} and $v_{A_{14}}$ have the same stalk $\{(b,d),(a,c)\}$. The stalks at the edges of G are determined by the stalks at the vertices. In this example, the stalks at the vertices or edges that correspond to 2D strata are determined uniquely by f, but the stalks at the vertices and edges that correspond to 0D or 1D strata depend on our choice of intrinsic indexing.

3. **The morphisms:** There are only three distinct nonidentity morphisms. The first two are

$$\begin{aligned}
\mathcal{F}_{v_{A_{23}} \le e_{(A_{23},Q_2)}},\, \mathcal{F}_{v_{\mathbf{0}} \le e_{(\mathbf{0},Q_2)}} : \{(a,c),(b,d)\} &\to \{(a,d),(b,c)\} \\
(a,c) &\mapsto (b,c) \\
(b,d) &\mapsto (a,d)\,, \\
\mathcal{F}_{v_{A_{34}} \le e_{(A_{34},Q_4)}},\, \mathcal{F}_{v_{\mathbf{0}} \le e_{(\mathbf{0},Q_4)}} : \{(a,d),(b,c)\} &\to \{(a,c),(b,d)\} \\
(a,d) &\mapsto (a,c) \\
(b,c) &\mapsto (b,d)\,.
\end{aligned}$$

The third distinct nonidentity morphism is

$$\mathcal{F}_{v_{\mathbf{0}} \le e_{(\mathbf{0},Q_1)}} : \{(a,d),(b,c)\} \to \{(a,c),(b,d)\}\,.$$

As we move from $\mathbf{0}$ to Q_1, we swap the simplex indices of a and b and we swap the simplex indices of c and d (in the simplex indexing induced by f). The morphism is not canonical because the bijection $\phi^{\mathrm{idx}_{\mathbf{0}},\mathrm{idx}_{Q_1}}$ depends on whether one first swaps a and b or one first swaps c and d. Therefore, we may define either

$$\begin{aligned}
\mathcal{F}_{v_{\mathbf{0}} \le e_{(\mathbf{0},Q_1)}} : (a,d) &\mapsto (a,c)\,, \\
(b,c) &\mapsto (b,d)
\end{aligned}$$

or

$$\mathcal{F}_{v_{\mathbf{0}} \leq e_{(\mathbf{0},Q_1)}} : (a,d) \mapsto (b,d)\,,$$
$$(b,c) \mapsto (a,c)\,.$$

Both choices results in a compatible cellular sheaf.

3.5.3 Sections of the cellular sheaf

Let $\mathcal{F}$ be any compatible cellular sheaf for a fibered filtration $f : \mathcal{K} \times B \to \mathbb{R}$. We write

$$\mathcal{F}_{v_\beta \leq e_{\beta,\alpha}} = \left(\mathcal{F}^b_{v_\beta \leq e_{\beta,\alpha}}, \mathcal{F}^d_{v_\beta \leq e_{\beta,\alpha}}\right), \tag{3.17}$$

where $\mathcal{F}^b_{v_\beta \leq e_{\beta,\alpha}} : S_\beta \to S_\alpha$ maps a pair $(\sigma_b, \sigma_d) \in S_\beta$ to the birth simplex of $\mathcal{F}_{v_\beta \leq e_{\beta,\alpha}}((\sigma_b, \sigma_d))$ and $\mathcal{F}^d_{v_\beta \leq e_{\beta,\alpha}} : S_\beta \to S_\alpha$ maps $(\sigma_b, \sigma_d) \in S_\beta$ to the death simplex of $\mathcal{F}_{v_\beta \leq e_{\beta,\alpha}}((\sigma_b, \sigma_d))$.

In this subsection, we show that one can view sections of $\mathcal{F}$ as sections of the associated PD bundle.

Lemma 3.5.3. Let Y_β be a face of Y_α. For any point $p \in Y_\beta$ and any pair (σ_b, σ_d) in $\mathcal{F}(v_\beta)$, we have

$$\Big(f(\sigma_b, p), f(\sigma_d, p)\Big) = \Big(f(\mathcal{F}^b_{v_\beta \leq e_{\beta,\alpha}}((\sigma_b, \sigma_d)), p)\,, f(\mathcal{F}^d_{v_\beta \leq e_{\beta,\alpha}}((\sigma_b, \sigma_d)), p)\Big)\,, \tag{3.18}$$

where $\mathcal{F}^b$ and $\mathcal{F}^d$ are defined as in Eq. 3.17.

Proof. If the simplex orderings in Y_α and Y_β differ only by a transposition of simplices (σ, τ) with consecutive indices in the orderings, then $\mathcal{F}_{v_\beta \leq e_{\beta,\alpha}}$ is either the identity map or the map that swaps σ and τ in the pairs that contain them. In either case, Equation 3.18 holds because $f(\sigma, p) = f(\tau, p)$ for all $p \in Q$. Equation 3.18 holds in general because $\mathcal{F}_{v_\beta \leq e_{\beta,\alpha}}$ is defined as the composition of such maps. □

Proposition 3.5.4. Let z_0 be a non-diagonal point in $PD_q(f_{p_0})$ for some $p_0 \in B$, let (σ_b, σ_d) be the (birth, death) simplex pair such that $z_0 = (f(\sigma_b, p_0), f(\sigma_d, p_0))$, and let Y_0 be the

stratum that contains p_0. Suppose that $\mathcal{F}$ is a compatible cellular sheaf, and let v_0 be the vertex in the graph G (see Definition 3.5.1) that is associated with Y_0. If there is a global section $\overline{s}$ of the cellular sheaf $\mathcal{F}$ such that $\overline{s}(v_0) = (\sigma_b, \sigma_d)$, then there is a global section s of the PD bundle such that $s(p_0) = z_0$.

Proof. Let $\overline{s}$ be a global section of the cellular sheaf $\mathcal{F}$ such that $\overline{s}(v_0) = (\sigma_b, \sigma_d)$. For every stratum Y_α, we write

$$\overline{s}(v_\alpha) = (\overline{s}_b(v_\alpha), \overline{s}_d(v_\alpha)) \,,$$

where $\overline{s}_b(v_\alpha)$ is the birth simplex of $\overline{s}(v_\alpha)$ and $\overline{s}_d(v_\alpha)$ is the death simplex of $\overline{s}(v_\alpha)$. Let $Y : B \to \{Y_\alpha\}$ be the function that maps $p \in B$ to the unique stratum Y_α that contains it.

We define $s : B \to E$ to be the function

$$s(p) := \Big(p, f(\overline{s}_b(v_{Y(p)}), p), f(\overline{s}_d(v_{Y(p)}), p)\Big) \,.$$

To show that $s : B \to E$ is a global section of the PD bundle, it remains to show that it is continuous. The function $s|_{Y_\alpha}$ is continuous for all strata Y_α because $f(\sigma, \cdot)$ is continuous for all simplices $\sigma \in \mathcal{K}$. Therefore, it suffices to show that $s|_{\overline{Y_\alpha}}$ is continuous on each face Y_β of Y_α. Because $\overline{s}$ is a section of the cellular sheaf,

$$\overline{s}(v_\alpha) = \mathcal{F}_{v_\beta \leq e_{\beta,\alpha}}(\overline{s}(v_\beta)) \,.$$

By Lemma 3.5.3,

$$\Big(f(\overline{s}_b(v_\beta), p), f(\overline{s}_d(v_\beta), p)\Big) = \Big(f(\mathcal{F}^b_{v_\beta \leq e_{\beta,\alpha}}(\overline{s}(v_\beta)), p), f(\mathcal{F}^d_{v_\beta \leq e_{\beta,\alpha}}(\overline{s}(v_\beta)), p)\Big)$$

for all points $p \in Y_\beta$. Therefore,

$$\Big(f(\overline{s}_b(v_\beta), p), f(\overline{s}_d(v_\beta), p)\Big) = \Big(f(\overline{s}_b(v_\alpha), p), f(\overline{s}_d(v_\alpha), p)\Big)$$

for all $p \in Y_\beta$, which completes the proof. □

3.6 Conclusions

3.6.1 Summary

In this chapter, I introduced the concept of a persistence-diagram (PD) bundle, a framework that can be used to study the persistent homology of a fibered filtration function (i.e., set of filtrations parameterized by an arbitrary "base space" B). Special cases of PD bundles include vineyards [CEM06], the persistent homology transform (PHT) [TMB14], the fibered barcode of a multiparameter persistence module [And13], and the barcode-decorated merge tree [CHM22].

In Theorem 3.3.15, I proved that if B is a smooth compact manifold, then for generic fibered filtrations, B is stratified so that the simplex ordering is constant within each stratum. When such a stratification exists, the PD bundle is determined by the PDs at a locally finite (or finite, if B is compact) subset of points in B. In Proposition 3.3.5, I showed that every piecewise-linear PD bundle has such a stratification into polyhedra. This polyhedral stratification is utilized in Chapter 4 ([Hic22a]) in an algorithm for computing piecewise-linear PD bundles.

I showed that, unlike vineyards, which PD bundles generalize, not every local section of a PD bundle can be extended to a global section (see Proposition 3.4.3). The implication of this result is that PD bundles do not necessarily have individual, separated "vines" in the way that vineyards do.

Lastly, I introduced a cellular sheaf that is compatible with a given PD bundle. In Proposition 3.5.4, I proved that one can determine whether a local section can be extended to a global section by determining whether or not there is an associated global section of a compatible cellular sheaf. A compatible cellular sheaf is a discrete mathematical data structure for summarizing the data in a PD bundle.

3.6.2 Discussion

For a given fibered filtration function f with a stratification as in Theorem 3.3.15, I defined a compatible cellular sheaf $\mathcal{F}$ over a graph G. It is tempting to instead define an associated cellular sheaf directly on the stratification of B. In particular, when f is piecewise linear, the strata are polyhedra, so the stratification is guaranteed to be a cellular decomposition. One could certainly define stalks $\mathcal{F}(Y_\alpha)$ and functions $\mathcal{F}(Y_\beta) \to \mathcal{F}(Y_\beta)$ in the same way as in Definition 3.5.1. The problem is that $\mathcal{F}$ would not necessarily satisfy the composition condition (see Equation 2.5). For instance, this issue occurs in Example 3.5.2 for the same reason that the morphism $\mathcal{F}_{v_\mathbf{0} \leq e_{(\mathbf{0},Q_1)}}$ in the example is not canonical (see the discussion in Example 3.5.2).

Additionally, I note that one could have defined a compatible cellular cosheaf rather than a sheaf.

3.6.3 Future research

I conclude with some questions and proposals for future work:

- What are the obstructions to the existence of a global section that extends a given local section?

  I conjecture that if $B \setminus B^*$ is contractible, where B^* is the set of singularities (i.e., points $p_* \in B$ at which there is a point in $\mathrm{PD}(f_{p_*})$ with multiplicity greater than 1), then every local section can be extended to a global section. I discussed in Remark 3.4.5 that singularities are typical when $\dim(B) > 1$. If the conjecture is true, then global sections are typically not guaranteed to exist when either $\dim(B) > 1$ or B is not contractible.

- What algebraic or computational methods can we use to analyze global sections and

to compute obstructions to the existence of global sections?

It may help to consider the cellular-sheaf perspective from Section 3.5.2, which turns the question into a discrete problem that one can study computationally.

- PHT is a PD bundle over the base space $B = S^n$. Are there constructible sets $M \subseteq \mathbb{R}^{n+1}$ for which the associated PHT has local sections that cannot be extended to global sections? What is the generic situation?

  In Proposition 3.4.3, we constructed a fibered filtration f with $B = \mathbb{R}^2$ that has no global sections. Additionally, in Remark 3.4.4, we noted that f can be restricted to $B = S^1$ to construct a PD bundle over S^1 that has no global sections. However, we do not know if this PD bundle is realized by the PHT of some constructible set $M \subseteq \mathbb{R}^2$.

- Arya et al. [ACM22] showed that the PHT of a constructible set M can be calculated by "gluing together" the PHT of smaller, simpler subsets of M. Can one generalize these results to all PD bundles?

- When are PD bundles "stable"?

  For example, vineyards are stable for generic 1-parameter filtrations, even though not every vineyard is stable (see Appendix 3.A for an example). Does an analogous result hold for generic fibered filtration functions over any base space B?

- It will also be interesting to study real-world applications of PD bundles, such as the examples that we mentioned in Section 3.2.1.

APPENDIX

3.A Vineyard instability

For any $\epsilon > 0$, we construct two 1-parameter filtration functions f_ϵ^+, f_ϵ^- that are ϵ perturbations of each other (that is, $|f_\epsilon^+(\sigma, p) - f_\epsilon^-(\sigma, p)| < \epsilon$ for all simplices $\sigma \in \mathcal{K}$ and all points $p \in B$) but such that for any bijection between the vines in the respective vineyards, not all of the matched vines are close to each other. In fact, we can define f_ϵ^+ and f_ϵ^- so that their vines are arbitrarily far apart.

We construct our example by restricting the filtration function from Section 3.4 to certain paths through $\mathbb{R}^2$. Let $\mathcal{K}$ and $f : \mathcal{K} \times \mathbb{R}^2 \to \mathbb{R}$ be defined as in the proof of Prop. 3.4.3. (See Figure 3.2b.) Because f is continuous, we have that for any $\epsilon > 0$, there is a $\delta > 0$ such that $|f(\sigma, p) - f(\sigma, \mathbf{0})| < \epsilon/2$ when $\|p\| < \sqrt{2}\delta$ and σ is any simplex in $\mathcal{K}$. We define the paths

$$\gamma_\epsilon^+(t) := \begin{cases} (t, t)\,, & |t| \geq \delta \\ (-\delta, \delta + 2t)\,, & -\delta < t < 0 \\ (-\delta + 2t, \delta)\,, & 0 \leq t < \delta\,, \end{cases}$$

$$\gamma_\epsilon^-(t) := \begin{cases} (t, t)\,, & |t| \geq \delta \\ (\delta + 2t, -\delta)\,, & -\delta < t < 0 \\ (\delta, -\delta + 2t)\,, & 0 \leq t < \delta\,. \end{cases}$$

Let $f_\epsilon^\pm : \mathcal{K} \times \mathbb{R} \to \mathbb{R}$ be the 1-parameter filtration function defined by $f^\pm(\sigma, t) := f(\sigma, \gamma_\epsilon^\pm(t))$. By construction, the filtrations f_ϵ^+ and f_ϵ^- are ϵ-perturbations of each other.

Let V^+ and V^- be the vineyards for f_ϵ^+ and f_ϵ^-, respectively, for the 1st degree PH. The

vineyards $V^{\pm}$ each have two vines $v_1^{\pm}$, $v_2^{\pm}$, which are paths $v_i^{\pm} : \mathbb{R} \to \mathbb{R}^3$. The vines are

$$v_1^+(t) = \begin{cases} (f(a,\gamma^+(t)), f(d,\gamma^+(t)))\,, & t \le -\delta/2 \\ f(b,\gamma^+(t)), f(d,\gamma^+(t)))\,, & t > -\delta/2\,, \end{cases}$$

$$v_2^+(t) = \begin{cases} (f(b,\gamma^+(t)), f(c,\gamma^+(t)))\,, & t \le -\delta/2 \\ f(a,\gamma^+(t)), f(c,\gamma^+(t)))\,, & t > -\delta/2\,, \end{cases}$$

$$v_1^-(t) = \begin{cases} (f(a,\gamma^-(t)), f(d,\gamma^-(t)))\,, & t \le -\delta/2 \\ f(a,\gamma^-(t)), f(c,\gamma^-(t)))\,, & t > -\delta/2\,, \end{cases}$$

$$v_2^-(t) = \begin{cases} (f(b,\gamma^-(t)), f(c,\gamma^-(t)))\,, & t \le -\delta/2 \\ f(b,\gamma^-(t)), f(d,\gamma^-(t)))\,, & t > -\delta/2\,. \end{cases}$$

There is no bijection $\phi : \{1,2\} \to \{1,2\}$ such that v_1^+ and v_2^+ are "close" to $v_{\phi(1)}^-$ and $v_{\phi(2)}^-$, respectively. This is because

$$\begin{aligned}
\left\|v_1^+(t) - v_1^-(t)\right\|^2 &= |f(b,(t,t)) - f(a,(t,t))|^2 + |f(d,(t,t))) - f(c,(t,t))|^2\,, && t > \delta/2\,, \\
\left\|v_1^+(t) - v_2^-(t)\right\|^2 &= |f(b,(t,t)) - f(a,(t,t))|^2 + |f(d,(t,t))) - f(c,(t,t))|^2\,, && t \le -\delta/2\,, \\
\left\|v_2^+(t) - v_2^-(t)\right\|^2 &= |f(b,(t,t)) - f(a,(t,t))|^2 + |f(d,(t,t))) - f(c,(t,t))|^2\,, && t > \delta/2\,, \\
\left\|v_2^+(t) - v_1^-(t)\right\|^2 &= |f(b,(t,t)) - f(a,(t,t))|^2 + |f(d,(t,t))) - f(c,(t,t))|^2\,, && t \le -\delta/2
\end{aligned}$$

and we can define f so that $|f(b,(t,t)) - f(a,(t,t)|$ and $|f(d,(t,t)) - f(c,(t,t))|$ are arbitrarily large for $t \neq 0$.

3.B Technical Details of Section 3.3

All notation is defined as in Section 3.3.

The following series of lemmas is used to prove Lemma 3.3.12, which shows that for almost every $\boldsymbol{a} \in A$, the tangent space of the intersection of sets $I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r})$ is equal to the

intersection of their tangent spaces.

Lemma 3.B.1. For almost every $\boldsymbol{a} \in A$, we have

$$T_p\Big(\bigcap_{j\in J} M_{\boldsymbol{a},j}\Big) = \bigcap_{j\in J} T_p(M_{\boldsymbol{a},j})$$

for all $J \subseteq \{1,\dots,N\}$ and all $p \in \bigcap_{j\in J} M_{\boldsymbol{a},j}$.

Proof. Because there are finitely many subsets of $\{1,\dots,N\}$, it suffices to show that for a given $J \subseteq \{1,\dots,N\}$, we have $T_p\Big(\bigcap_{j\in J} M_{\boldsymbol{a},j}\Big) = \bigcap_{j\in J} T_p(M_{\boldsymbol{a},j})$ for all $p \in \bigcap_{j\in J} M_{\boldsymbol{a},j}$ for almost every $\boldsymbol{a} \in A$. Let $\{j_i\}_{i=1}^k$ be the elements of J, where $j_i < j_{i+1}$ for all i. Because transverse intersections are generic, we have $M_{\boldsymbol{a},j_i} \pitchfork (M_{\boldsymbol{a},j_1} \cap \cdots \cap M_{\boldsymbol{a},j_{i-1}})$ for every i for almost every $\boldsymbol{a} \in A$. For such an $\boldsymbol{a} \in A$, we have

$$T_p\Big(\bigcap_{j\in J} M_{\boldsymbol{a},j}\Big) = T_p(M_{\boldsymbol{a},j_k}) \cap T_p\Big(\bigcap_{i=1}^{k-1} M_{\boldsymbol{a},j_i}\Big)$$

because $M_{\boldsymbol{a},j_k} \pitchfork (M_{\boldsymbol{a},j_1} \cap \cdots \cap M_{\boldsymbol{a},j_{k-1}})$. Therefore,

$$T_p\Big(\bigcap_{j\in J} M_{\boldsymbol{a},j}\Big) = \bigcap_{i=1}^{k} T_p(M_{\boldsymbol{a},j_i})$$

by induction on i. □

Lemma 3.B.2. Let $\boldsymbol{a} \in A$, and let $\{(i_r, j_r)\}_{r=1}^m$ be a set of index pairs such that $\{i_r, j_r\} \neq \{i_s, j_s\}$ if $r \neq s$. If B is a compact manifold, then there is a finite open cover $\{U_k\}_{k=1}^K$ and a disjoint partition $\bigcup_\ell J_{\ell,k} = \{1,\dots,m\}$ for each k such that

$$\{i_r, j_r \mid r \in J_{\ell_1,k}\} \cap \{i_r, j_r \mid r \in J_{\ell_2,k}\} = \emptyset$$

if $\ell_1 \neq \ell_2$ and

$$\pi\Big(\bigcap_{r\in J_{\ell,k}} (M_{\boldsymbol{a},i_r} \cap M_{\boldsymbol{a},j_r})\Big) \cap U_k = \bigcap_{r\in J_{\ell,k}} I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r}) \cap U_k$$

for all ℓ, where π is the projection $\pi : B \times \mathbb{R} \to B$.

Proof. Suppose that $y \in \bigcap_{r=1}^{m} I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r})$. Let $J_\ell^0 := \{\ell\}$ be an initial disjoint partition of $\{1, \dots, m\}$. By definition, $\pi(M_{\boldsymbol{a},i_\ell} \cap M_{\boldsymbol{a},j_\ell}) = I_{\boldsymbol{a}}(\sigma_{i_\ell}, \sigma_{j_\ell})$. If $i_{\ell_1} = i_{\ell_2}$ for some $\ell_1 \neq \ell_2$, then $f(\sigma_{j_{\ell_1}}, y) = f(\sigma_{i_{\ell_1}}, y) = f(\sigma_{j_{\ell_2}}, y)$, so $y \in \pi(M_{\boldsymbol{a},i_{\ell_1}} \cap M_{\boldsymbol{a},j_{\ell_1}} \cap M_{\boldsymbol{a},i_{\ell_2}} \cap M_{\boldsymbol{a},j_{\ell_2}})$. We combine $J_{\ell_1}^0$ and $J_{\ell_2}^0$ into a single subset of the partition, and we iterate until we obtain a disjoint partition $\{J_{\ell,y}\}_\ell$ of $\{1, \dots, m\}$ such that

$$\{i_r, j_r \mid r \in J_{\ell_1,y}\} \cap \{i_r, j_r \mid r \in J_{\ell_2,y}\} = \emptyset$$

if $\ell_1 \neq \ell_2$ and

$$y \in \pi\Big(\bigcap_{r \in J_{\ell,y}} (M_{\boldsymbol{a},i_r} \cap M_{\boldsymbol{a},j_r}) \Big)$$

for all ℓ. Therefore, for each ℓ, there is a neighborhood $U_{\ell,y}$ such that

$$\pi\Big(\bigcap_{r \in J_{\ell,y}} (M_{\boldsymbol{a},i_r} \cap M_{\boldsymbol{a},j_r}) \Big) \cap U_{\ell,y} = \bigcap_{r \in J_{\ell,y}} I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r}) \cap U_{\ell,k}\,.$$

Now, we set $U_y := \bigcap_\ell U_{\ell,y}$. Because B is compact, there is a finite open cover $\{U_k\}_{k=1}^K$ that has the desired properties by construction. □

Lemma 3.B.3. Let $\{(i_r, j_r)\}_{r=1}^m$ be a set of index pairs such that $\{i_r, j_r\} \neq \{i_s, j_s\}$ if $r \neq s$. For almost every $\boldsymbol{a} \in A$, we have that if

1. $\bigcup_\ell J_\ell = \{1, \dots, m\}$ is a disjoint partition such that

$$\{i_r, j_r \mid r \in J_{\ell_1,k}\} \cap \{i_r, j_r \mid r \in J_{\ell_2,k}\} = \emptyset$$

   for $\ell_1 \neq \ell_2$ and

2. U is an open set in B such that

$$\pi\Big(\bigcap_{r \in J_\ell} (M_{\boldsymbol{a},i_r} \cap M_{\boldsymbol{a},j_r}) \Big) \cap U = \bigcap_{r \in J_\ell} I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r}) \cap U\,, \tag{3.19}$$

   for all ℓ, where π is the projection $\pi : B \times \mathbb{R} \to B$,

then

1. the set $\bigcap_{r\in J'} I_{\boldsymbol{a}}(\sigma_{i_r},\sigma_{j_r})$ is a manifold for every $J'\subseteq\{1,\dots,m\}$ and

2. we have
$$T_y\Big(\bigcap_{\ell}\bigcap_{r\in J_\ell} I_{\boldsymbol{a}}(\sigma_{i_r},\sigma_{j_r})\Big)=\bigcap_{\ell} T_y\Big(\bigcap_{r\in J_\ell} I_{\boldsymbol{a}}(\sigma_{i_r},\sigma_{j_r})\Big) \tag{3.20}$$
for every $y\in\bigcap_{r=1}^{m} I_{\boldsymbol{a}}(\sigma_{i_r},\sigma_{j_r})\cap U$.

Proof. It suffices to show that
$$\Big(\bigcap_{r\in J_\ell} I_{\boldsymbol{a}}(\sigma_{i_r},\sigma_{j_r})\cap U\Big)\pitchfork\Big(\bigcap_{\ell'<\ell}\bigcap_{r\in J_\ell} I_{\boldsymbol{a}}(\sigma_{i_r},\sigma_{j_r})\cap U\Big) \tag{3.21}$$
for all ℓ and almost every $\boldsymbol{a}\in A$.

Informally, what we show first is that at almost every $\boldsymbol{a}\in A$, perturbations of $\boldsymbol{a}$ produce perturbations of $\bigcap_{r\in J_\ell} I_{\boldsymbol{a}}(\sigma_{i_r},\sigma_{j_r})\cap U$ for each ℓ. By Lemmas 3.3.8 and 3.3.9, we may assume that $\bigcap_{r\in J'} I_{\boldsymbol{a}}(\sigma_{i_r},\sigma_{j_r})$ is a manifold for every $J'\subseteq\{1,\dots,m\}$. By Equation 3.22, we may assume without loss of generality that there is a sequence $k_1<\cdots<k_c$ such that $j_1=k_1$ and $i_r=j_{r-1}$ for all r and $j_{r+1}=i_r$ for all r. In other words, we may assume that $\bigcap_{r\in J_\ell} I_{\boldsymbol{a}}(\sigma_{i_r},\sigma_{j_r})$ is of the form
$$I_{\boldsymbol{a}}(\sigma_{k_c},\sigma_{k_{c-1}})\cap\cdots\cap I_{\boldsymbol{a}}(\sigma_{k_3},\sigma_{k_2})\cap I_{\boldsymbol{a}}(\sigma_{k_2},\sigma_{k_1})$$
for all $\boldsymbol{a}$. The idea is that because the intersection lifts to an intersection of the corresponding manifolds (see Equation 3.22) we can pair them up however we like.

Define the function $g_i:B\to\mathbb{R}$ by
$$g_i(p):=f(\sigma_{k_i},p)-f(\sigma_{k_{i-1}},p)\,.$$
We will repeatedly use the following fact: If $g:B\to\mathbb{R}$ is a smooth map and $y\in\mathbb{R}$ is a regular value with preimage Z_y, then y is a regular value of $g|_Z$ for a submanifold $Z\subseteq X$ if $Z\pitchfork Z_y$. For almost every $\boldsymbol{a}\in A$, the quantity $a_{k_i}-a_{k_{i-1}}$ is a regular value of g_i for all i,and the set of regular values is open. By the same argument as in the proof of Lemma 3.3.9, we

have

$$\Big(I_{\boldsymbol{a}}(\sigma_{k_c},\sigma_{k_{c-1}})\cap U\Big)\pitchfork\Big(\bigcap_{i=2}^{c-1}I_{\boldsymbol{a}}(\sigma_{k_i},\sigma_{k_{i-1}})\cap U\Big)$$

for almost every $\boldsymbol{a}\in A$. Therefore, $(a_{k_c}-a_{k_{c-1}})$ is a regular value of $g_k|_{\bigcap_{i=2}^{c-1}I_{\boldsymbol{a}}(\sigma_{k_i},\sigma_{k_{i-1}})\cap U}$. Additionally, for $\boldsymbol{\epsilon}\in\mathbb{R}^N$ such that ϵ_{k_c} and $\epsilon_{k_{c-1}}$ are sufficiently small, there are no critical values between $(a_{k_c}-a_{k_{c-1}})$ and $(a_{k_c}-a_{k_{c-1}}+\epsilon_{k_c}-\epsilon_{k_{c-1}})$. Because there are no critical values, the set $\bigcap_{r\in J_\ell}I_{\boldsymbol{a}}(\sigma_{i_r},\sigma_{j_r})\cap U$ (which is the $(a_{k_c}-a_{k_{c-1}})$-level set of $g_k|_{\bigcap_{i=2}^{c-1}I_{\boldsymbol{a}}(\sigma_{k_i},\sigma_{k_{i-1}})\cap U}$) is a submanifold of B that is diffeomorphic to $I_{\boldsymbol{a}+\boldsymbol{\epsilon}}(\sigma_{k_c},\sigma_{k_{c-1}})\cap\Big(\bigcap_{i=2}^{c-1}I_{\boldsymbol{a}}(\sigma_{k_i},\sigma_{k_{i-1}})\Big)\cap U$ (which is the $(a_{k_c}-a_{k_{c-1}}+\epsilon_{k_c}-\epsilon_{k_{c-1}})$-level set of $g_k|_{\bigcap_{i=2}^{c-1}I_{\boldsymbol{a}}(\sigma_{k_i},\sigma_{k_{i-1}})\cap U}$), and these submanifolds are smoothly parameterized by $\epsilon_{k_c},\epsilon_{k_{c-1}}$.

Now consider any $i_*\in\{2,\dots,c-1\}$. Because $(a_{k_{i_*}}-a_{k_{i_*-1}})$ is a regular value of g_{i_*} and the set of regular values is open, there are no critical values between $(a_{k_{i_*}}-a_{k_{i_*-1}})$ and $(a_{k_{i_*}}-a_{k_{i_*-1}}-\epsilon_{k_{i_*-1}})$ for sufficiently small $\epsilon_{k_{i_*-1}}$. Therefore, for sufficiently small $\epsilon_{k_{i_*-1}}$, the $(a_{k_{i_*}}-a_{k_{i_*-1}}-\epsilon_{k_{i_*-1}})$-level set of g_{i_*} is a submanifold of B that is diffeomorphic to the $(a_{k_{i_*}}-a_{k_{i_*-1}})$-level set, and these submanifolds are smoothly parameterized by (sufficiently small) $\epsilon_{k_{i_*-1}}$. Because transverse intersections are generic,

$$\Big(\bigcap_{i=i_*+1}^{c}I_{\boldsymbol{a}+\boldsymbol{\epsilon}}(\sigma_{k_i},\sigma_{k_{i-1}})\Big)\cap\Big(\bigcap_{i=2}^{i_*-1}I_{\boldsymbol{a}+\boldsymbol{\epsilon}}(\sigma_{k_i},\sigma_{k_{i-1}})\Big)\cap U$$

is transverse to the $(a_{k_{i_*}}-a_{k_{i_*-1}}-\epsilon_{k_{i_*-1}})$-level set of g_{i_*} for almost every (sufficiently small) $\epsilon_{k_{i_*-1}}$. Additionally, if the intersection is transverse, it is transverse for an open neighborhood of $\epsilon_{k_{i_*-1}}$. Therefore, we can assume without loss of generality that this intersection is transverse at $\epsilon_{k_{i_*-1}}=0$ (if not, we can perturb $a_{k_{i_*-1}}$ so that it is) and for all sufficiently small $\epsilon_{k_{i_*-1}}$. This implies that $(a_{k_{i_*}}-a_{k_{i_*-1}}-\epsilon_{k_{i_*-1}})$ is also a regular value of g_{i_*} restricted to $\Big(\bigcap_{i=i_*+1}^{c}I_{\boldsymbol{a}+\boldsymbol{\epsilon}}(\sigma_{k_i},\sigma_{k_{i-1}})\Big)\cap\Big(\bigcap_{i=2}^{i_*-1}I_{\boldsymbol{a}+\boldsymbol{\epsilon}}(\sigma_{k_i},\sigma_{k_{i-1}})\Big)\cap U$. For sufficiently small $\epsilon_{k_{i_*}}$, there are no critical values between $(a_{k_{i_*}}-a_{k_{i_*-1}}-\epsilon_{k_{i_*-1}})$ and $(a_{k_{i_*}}-a_{k_{i_*-1}}+\epsilon_{k_{i_*}}-\epsilon_{k_{i_*-1}})$. Therefore, for sufficiently small $\epsilon_{k_{i_*}}$, $\epsilon_{k_{i_*-1}}$, we have that

$$\Big(\bigcap_{i=i_*+1}^{c}I_{\boldsymbol{a}+\boldsymbol{\epsilon}}(\sigma_{k_i},\sigma_{k_{i-1}})\Big)\cap\Big(\bigcap_{i=2}^{i_*}I_{\boldsymbol{a}+\boldsymbol{\epsilon}}(\sigma_{k_i},\sigma_{k_{i-1}})\Big)\cap U\,,$$

which is the $(a_{k_{i_*}} - a_{k_{i_*-1}})$-level set of g_{i_*} restricted to

$$\Big(\bigcap_{i=i_*+1}^{c} I_{\boldsymbol{a}+\boldsymbol{\epsilon}}(\sigma_{k_i}, \sigma_{k_{i-1}}) \Big) \cap \Big(\bigcap_{i=2}^{i_*-1} I_{\boldsymbol{a}+\boldsymbol{\epsilon}}(\sigma_{k_i}, \sigma_{k_{i-1}}) \Big) \cap U \,,$$

is a submanifold of B that is diffeomorphic to

$$\Big(\bigcap_{i=i_*}^{c} I_{\boldsymbol{a}+\boldsymbol{\epsilon}}(\sigma_{k_i}, \sigma_{k_{i-1}}) \Big) \cap \Big(\bigcap_{i=2}^{i_*-1} I_{\boldsymbol{a}+\boldsymbol{\epsilon}}(\sigma_{k_i}, \sigma_{k_{i-1}}) \Big) \cap U \,,$$

which is the $(a_{k_{i_*}} - a_{k_{i_*-1}} + \epsilon_{k_{i_*}} - \epsilon_{k_{i_*-1}})$-level set of g_{i_*} restricted to $\Big(\bigcap_{i=i_*+1}^{c} I_{\boldsymbol{a}+\boldsymbol{\epsilon}}(\sigma_{k_i}, \sigma_{k_{i-1}}) \Big) \cap \Big(\bigcap_{i=2}^{i_*-1} I_{\boldsymbol{a}+\boldsymbol{\epsilon}}(\sigma_{k_i}, \sigma_{k_{i-1}}) \Big) \cap U$. These submanifolds are smoothly parameterized by $\epsilon_{k_{i_*}}$ and $\epsilon_{k_{i_*+1}}$.

By induction on i_*, we conclude that there is a set $A' \subseteq A$ such that $A \setminus A'$ has measure zero and such that for all $\boldsymbol{a} \in A'$, we have that

1. the set $\bigcap_{r \in J_\ell} I_{\boldsymbol{a}+\boldsymbol{\epsilon}}(\sigma_{i_r}, \sigma_{j_r}) \cap U$ is a submanifold of B that is diffeomorphic to

$$\bigcap_{r \in J_\ell} I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r}) \cap U$$

   for sufficiently small $\boldsymbol{\epsilon} \in \mathbb{R}^N$, and

2. these submanifolds are smoothly parameterized by $\boldsymbol{\epsilon}$.

Let $\boldsymbol{a} \in A'$. We showed above that for sufficiently small $\boldsymbol{\epsilon} \in \mathbb{R}^N$, the set of manifolds $\bigcap_{r \in \ell} I_{\boldsymbol{a}+\boldsymbol{\epsilon}}(\sigma_{i_r}, \sigma_{j_r}) \cap U$ (parameterized by $\boldsymbol{\epsilon}$) is a smoothly parameterized family of embeddings of $\bigcap_{r \in \ell} I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r}) \cap U$ into $U \subseteq B$. Varying $\{\epsilon_r \mid r \in J_\ell\}$ (while holding ϵ_r constant for $r \notin J_\ell$) produces a smoothly parameterized family of embeddings of $\bigcap_{r \in \ell} I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r}) \cap U$ while holding $\Big(\bigcap_{\ell' < \ell} \bigcap_{r \in J_\ell} I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r}) \cap U \Big)$ constant. Therefore, because transverse intersections are generic,

$$\Big(\bigcap_{r \in J_\ell} I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r}) \cap U \Big) \pitchfork \Big(\bigcap_{\ell' < \ell} \bigcap_{r \in J_\ell} I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r}) \cap U \Big)$$

for all ℓ for almost every $\boldsymbol{\epsilon}$ in a neighborhood of $\mathbf{0} \in \mathbb{R}^N$. This proves Eq. 3.21, which completes the proof. □

Lemma 3.B.4. Let $\{(i_r, j_r)\}_{r=1}^m$ be a set of index pairs such that $\{i_r, j_r\} \neq \{i_s, j_s\}$ if $r \neq s$. For almost every $\boldsymbol{a} \in A$, we have that if U is an open set in B such that

$$\pi\Big(\bigcap_{r\in J_\ell}(M_{\boldsymbol{a},i_r} \cap M_{\boldsymbol{a},j_r})\Big) \cap U = \bigcap_{r\in J_\ell} I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r}) \cap U \tag{3.22}$$

for all ℓ, where π is the projection $\pi : B \times \mathbb{R} \to B$, then $\bigcap_{r\in J'} I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r})$ is a manifold for every $J' \subseteq \{1, \ldots, m\}$ and

$$T_y\Big(I_{\boldsymbol{a}}(\sigma_{i_1}, \sigma_{j_1}) \cap \cdots I_{\boldsymbol{a}}(\sigma_{i_m}, \sigma_{j_m})\Big) = \bigcap_{r=1}^m T_y\Big(I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r})\Big) \tag{3.23}$$

for all $y \in \bigcap_{r=1}^m I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r}) \cap U$.

Proof. By Lemmas 3.3.8 and 3.3.9, $\bigcap_{r\in J'} I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r})$ is a manifold for every $J' \subseteq \{1, \ldots, m\}$ for almost every $\boldsymbol{a} \in A$. We have

$$T_y\Big(I_{\boldsymbol{a}}(\sigma_{i_1}, \sigma_{j_1}) \cap \cdots \cap I_{\boldsymbol{a}}(\sigma_{i_m}, \sigma_{j_m})\Big) \subseteq \bigcap_{r=1}^m T_y\Big(I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r})\Big)$$

because $\bigcap_{r=1}^m I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r}) \subseteq I_{\boldsymbol{a}}(\sigma_{i_s}, \sigma_{j_s})$ for all s.

Let $\boldsymbol{v} \in \bigcap_{r=1}^m T_y\Big(I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r})\Big)$. Define $\pi_{[m]} := \pi|_{\bigcap_{r=1}^m (M_{\boldsymbol{a},i_r} \cap M_{\boldsymbol{a},j_r}) \cap \pi^{-1}(U)}$, and define $\pi_r := \pi|_{M_{\boldsymbol{a},i_r} \cap M_{\boldsymbol{a},j_r} \cap \pi^{-1}(U)}$ for each r. Each π_r is a diffeomorphism from $M_{\boldsymbol{a},i_r} \cap M_{\boldsymbol{a},j_r} \cap \pi^{-1}(U)$ to $I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r}) \cap U$, and $\pi_{[m]}$ is a diffeomorphism from $\bigcap_{r=1}^m (M_{\boldsymbol{a},i_r} \cap M_{\boldsymbol{a},j_r}) \cap \pi^{-1}(U)$ to $\bigcap_{r=1}^m I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r}) \cap U$. Let $\tilde{y} := \pi_{[m]}^{-1}$ (which exists because $\pi_{[m]}^{-1}$ is a diffeomorphism), and let $\tilde{v} := d\pi_{[m]}^{-1}(v)$ (which exists because $d\pi_{[m]}$ is an isomorphism). For all r, we have $\tilde{y} = \pi_r^{-1}(y)$ and $\tilde{v} := d\pi_r^{-1}(v)$. Therefore,

$$\tilde{v} \in \bigcap_{r=1}^m T_{\tilde{y}}\Big(M_{\boldsymbol{a},i_r} \cap M_{\boldsymbol{a},j_r}\Big).$$

By Lemma 3.B.1, we have $T_{\tilde{y}}\Big(\bigcap_{r=1}^m (M_{\boldsymbol{a},i_r} \cap M_{\boldsymbol{a},j_r})\Big) = \bigcap_{r=1}^m T_{\tilde{y}}(M_{\boldsymbol{a},i_r} \cap M_{\boldsymbol{a},j_r})$ for all $\tilde{y} \in \bigcap_{r=1}^m (M_{\boldsymbol{a},i_r} \cap M_{\boldsymbol{a},j_r})$ for almost every $\boldsymbol{a} \in A$, so

$$\tilde{v} \in T_{\tilde{y}}\Big(\bigcap_{r=1}^m (M_{\boldsymbol{a},i_r} \cap M_{\boldsymbol{a},j_r})\Big)$$

for almost every $\boldsymbol{a} \in A$. Therefore, $v = d\pi_{[m]}(\tilde{v})$ is in $T_y\Big(\pi_{[m]}\Big(\bigcap_{r=1}^m (M_{\boldsymbol{a},i_r} \cap M_{\boldsymbol{a},j_r})\Big)\Big) = T_y\Big(\bigcap_{r=1}^m I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r})\Big)$, which implies that

$$T_y\Big(\bigcap_{r=1}^m I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r})\Big) \supseteq \bigcap_{r=1}^m T_y\Big(I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r})\Big).$$

□

The following series of lemmas is used to prove Lemma 3.3.13, which shows that $\mathcal{Y}_{\boldsymbol{a}}$ is locally finite for almost every $\boldsymbol{a} \in A$, and Lemma 3.3.14, which shows that $\mathcal{Y}_{\boldsymbol{a}}$ satisfies the Axiom of the Frontier for almost every $\boldsymbol{a} \in A$.

Lemma 3.B.5. If $\boldsymbol{a} \in A$ is such that each $Y \in \mathcal{Y}_{\boldsymbol{a}}$ is a manifold, then for any strict partial order $\prec$ on $\mathcal{K}$, there is a unique subset $\mathcal{Y}_{\boldsymbol{a}}^{\prec} \subseteq \mathcal{Y}_{\boldsymbol{a}}$ such that $Z_{\boldsymbol{a}}^{\prec} = \bigcup_{Y \in \mathcal{Y}_{\boldsymbol{a}}^{\prec}} Y$.

Proof. Let $Y \in \mathcal{Y}_{\boldsymbol{a}}$ and suppose that $Y \cap Z_{\boldsymbol{a}}^{\prec} \neq \emptyset$. This implies that there is a point $p \in Y$ such that $\prec_{f_{\boldsymbol{a}}(\cdot,p)}$ is the same as $\prec$. By Lemma 3.3.11, the simplex ordering induced by f is constant in Y, so $Y \subseteq Z_{\boldsymbol{a}}^{\prec}$. □

Lemma 3.B.6. Let $\prec$ be a strict partial order on $\mathcal{K}$. Let $\boldsymbol{a} \in A$ be such that

1. every $Y \in \mathcal{Y}_{\boldsymbol{a}}$ is a manifold, where $\mathcal{Y}_{\boldsymbol{a}}$ is defined as in Eq. 3.11,
2. $M_{\boldsymbol{a},i} \pitchfork M_{\boldsymbol{a},j}$ for all $i \neq j$, where $M_{\boldsymbol{a},i}$ is defined as in Eq. 3.5,
3. $\bigcap_{r=1}^m I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r})$ is a manifold for all sets $\{(i_r, j_r\}_{r=1}^m$ of index pairs, and
4. we have
$$T_p\Big(\bigcap_{r=1}^m I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r})\Big) = \bigcap_{r=1}^m T_p\Big(I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r})\Big) \tag{3.24}$$
for all sets $\{(i_r, j_r\}_{r=1}^m$ of index pairs and all $p \in \bigcap_{r=1}^m I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r})$.

Let $\mathcal{Y}_{\boldsymbol{a}}^{\prec}$ be the unique subset of $\mathcal{Y}_{\boldsymbol{a}}$ such that $Z_{\boldsymbol{a}}^{\prec} = \bigcup_{Y \in \mathcal{Y}_{\boldsymbol{a}}^{\prec}} Y$, which exists by Lemma 3.B.5. Then every $p \in B$ has a neighborhood that intersects at most one set $Y \in \mathcal{Y}_{\boldsymbol{a}}^{\prec}$.

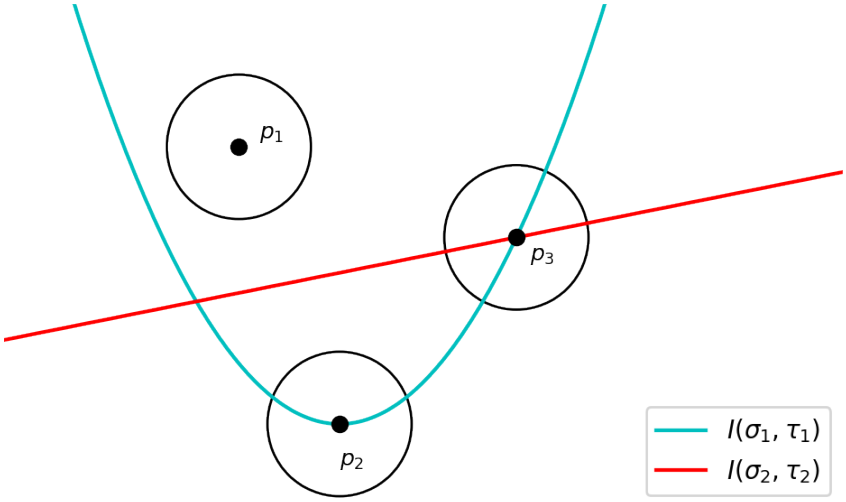


Figure 3.B.1: For each point p_i, we illustrate the idea behind the homeomorphism $\phi : U_i \to \mathcal{B}$ in the proof of Lemma 3.B.6, where U_i is a neighborhood of p_i and $\mathcal{B}$ is an open unit ball in $\mathbb{R}^2$. The base space is $B = \mathbb{R}^2$; we show only the sets $I(\sigma_1, \tau_1)$ and $I(\sigma_2, \tau_2)$, which are curves in the plane, for some simplices $\sigma_1, \sigma_2, \tau_1, \tau_2 \in \mathcal{K}$ and some fibered filtration function $f : \mathcal{K} \times B \to \mathbb{R}$.

Proof. Let $S(p) = \{(\sigma_i, \sigma_j) \mid p \in I_{\boldsymbol{a}}(\sigma_i, \sigma_j)\}$. There is a neighborhood U_0 of p such that $I(\sigma_i, \sigma_j) \cap U_0 \neq \emptyset$ if and only if $(\sigma_i, \sigma_j) \in S(p)$. In a neighborhood of p, each $I(\sigma_i, \sigma_j)$ is locally diffeomorphic (via the exponential map, for example) to $T_p(I_{\boldsymbol{a}}(\sigma_i, \sigma_j))$, which is an $(n-1)$-dimensional hyperplane. By Eq. 3.24, these local diffeomorphisms are compatible with each other, so there is a neighborhood U of p, a set $\{H(\sigma_i, \sigma_j)\}_{(\sigma_i,\sigma_j)\in S(p)}$ of hyperplanes in $\mathbb{R}^n$, and a homeomorphism $\phi : U \to \mathcal{B}$, where $\mathcal{B}$ is the open unit n-ball, such that

$$\phi\Big(\bigcap_{(\sigma_i,\sigma_j)\in S'(p)} I_{\boldsymbol{a}}(\sigma_i, \sigma_j) \cap U \Big) = \bigcap_{(\sigma_i,\sigma_j)\in S'(p)} H(\sigma_i, \sigma_j) \cap \mathcal{B}$$

for all $S'(p) \subseteq S(p)$. See Figure 3.B.1 for intuition, where we illustrate the neighborhood U for a few points $p \in B$. The hyperplanes induce a stratification of $\mathcal{B}$, with a set $\mathcal{Y}'$ of strata, such that for all $Y' \in \mathcal{Y}'$, we have $Y' = \phi(Y \cap U)$ for some $Y \in \mathcal{Y}_{\boldsymbol{a}}$. Because $M_{\boldsymbol{a},i} \pitchfork M_{\boldsymbol{a},j}$ for all $i \neq j$, we have that $\mathcal{B} \setminus H(\sigma_i, \sigma_j)$ is the disjoint union of open sets $W_1(\sigma_i, \sigma_j)$ and

$W_2(\sigma_i, \sigma_j)$ such that

$$f_{\boldsymbol{a}}(\sigma_i, p') < f_{\boldsymbol{a}}(\sigma_j, p') \qquad \text{for all } p' \in \phi^{-1}(W_1(\sigma_i, \sigma_j))\,,$$
$$f_{\boldsymbol{a}}(\sigma_j, p') < f_{\boldsymbol{a}}(\sigma_i, p') \qquad \text{for all } p' \in \phi^{-1}(W_2(\sigma_i, \sigma_j))$$

for all $(\sigma_i, \sigma_j) \in S(p)$. Suppose that u_1 and u_2 are points in U such that $\prec_{f_{\boldsymbol{a}}(\cdot,u_1)}$ and $\prec_{f_{\boldsymbol{a}}(\cdot,u_2)}$ are both the same as $\prec$. For each $(\sigma_i, \sigma_j) \in S(p)$, define the set

$$V(\sigma_i, \sigma_j) := \begin{cases} H(\sigma_i, \sigma_j)\,, & \sigma_i \not\prec \sigma_j \text{ and } \sigma_j \not\prec \sigma_i \\ W_1(\sigma_i, \sigma_j)\,, & \sigma_i \prec \sigma_j \\ W_2(\sigma_i, \sigma_j)\,, & \sigma_j \prec \sigma_i\,. \end{cases} \tag{3.25}$$

We define

$$V := \bigcap_{(\sigma_i,\sigma_j)\in S(p)} V(\sigma_i, \sigma_j) \neq \emptyset\,, \tag{3.26}$$

which is a stratum in $\mathcal{Y}'$. Therefore, there is a $Y \in \mathcal{Y}_{\boldsymbol{a}}$ such that $Y \cap U = \phi^{-1}(V)$, with $u_1, u_2 \in Y \cap U$. Therefore, Y is the only element of $\mathcal{Y}_{\boldsymbol{a}}^{\prec}$ that U intersects. □

Lemma 3.B.7. Let $\prec_0$ be a strict partial order on $\mathcal{K}$, and define $\mathcal{O}$ to be the set of strict partial orders $\prec$ such that

1. if $\sigma \prec_0 \tau$ and $\sigma \neq \tau$, then either we have $\sigma \prec \tau$ or we have $\sigma \not\prec \tau$ and $\tau \not\prec \sigma$,
2. if $\sigma \not\prec_0 \tau$ and $\tau \not\prec_0 \sigma$, then $\sigma \not\prec \tau$ and $\tau \not\prec \sigma$, and
3. the strict partial order $\prec$ is not the same as $\prec_0$.

If $\boldsymbol{a} \in A$ is such that

1. every $S \in \overline{E_{\boldsymbol{a}}^{n-\ell}}$ is an ℓ-dimensional smooth submanifold for every $\ell \in \{1, \ldots, n\}$, where n is the dimension of B,
2. $M_{\boldsymbol{a},i} \pitchfork M_{\boldsymbol{a},j}$ for all $i \neq j$,

3. the set $\bigcap_{r=1}^{m} I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r})$ is a manifold for all sets $\{(i_r, j_r)\}_{r=1}^{m}$ of index pairs, and

4. $T_y(\bigcap_{r=1}^{m} I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r}) = \bigcap_{r=1}^{m} T_y(I_{\boldsymbol{a}}(\sigma_{i_r}, \sigma_{j_r}))$ for all sets $\{(i_r, j_r)\}_{r=1}^{m}$ of index pairs,

then

$$\partial Z_{\boldsymbol{a}}^{\prec_0} = \bigcup_{\prec \text{ in } \mathcal{O}} Z_{\boldsymbol{a}}^{\prec} .$$

Proof. By Lemmas 3.3.10, every $Y \in \mathcal{Y}_{\boldsymbol{a}}$ is a manifold. By Lemmas 3.B.5 and 3.B.6, the sets $Z_{\boldsymbol{a}}^{\prec_0}$ and $Z_{\boldsymbol{a}}^{\prec}$ (for all $\prec$ in $\mathcal{O}$) are submanifolds of B.

Case 1: If $\dim(Z_{\boldsymbol{a}}^{\prec_0}) = 0$, then we must have

$$Z_{\boldsymbol{a}}^{\prec_0} = I_{\boldsymbol{a}}(\sigma_{i_1}, \sigma_{j_1}) \cap \cdots \cap I_{\boldsymbol{a}}(\sigma_{i_n}, \sigma_{j_n}) .$$

for some $I_{\boldsymbol{a}}(\sigma_{i_1}, \sigma_{j_1}) \cap \cdots \cap I_{\boldsymbol{a}}(\sigma_{i_n}, \sigma_{j_n}) \in \overline{E_{\boldsymbol{a}}^{n}}$. If $\prec$ is in $\mathcal{O}$, then there is another pair $(\sigma_{i_{n+1}}, \sigma_{j_{n+1}})$ of distinct simplices such that $\sigma_{i_{n+1}} \not\prec \sigma_{j_{n+1}}$ and $\sigma_{j_{n+1}} \not\prec \sigma_{i_{n+1}}$. Therefore,

$$Z_{\boldsymbol{a}}^{\prec} \subseteq I_{\boldsymbol{a}}(\sigma_{i_1}, \sigma_{j_1}) \cap \cdots \cap I_{\boldsymbol{a}}(\sigma_{i_n}, \sigma_{j_n}) \cap I_{\boldsymbol{a}}(\sigma_{i_{n+1}}, \sigma_{j_{n+1}}) ,$$

which is an element of $\overline{E_{\boldsymbol{a}}^{n+1}}$. By choice of $\boldsymbol{a}$, every $S \in \overline{E_{\boldsymbol{a}}^{n+1}}$ is empty, so $Z_{\boldsymbol{a}}^{\prec} = \emptyset$.

Case 2: If $\dim(Z_{\boldsymbol{a}}^{\prec_0}) \geq 1$, let $\prec$ be any strict partial order in $\mathcal{O}$. Let $p \in Z_{\boldsymbol{a}}^{\prec}$. Let $S(p) = \{(\sigma_i, \sigma_j) \mid p \in I_{\boldsymbol{a}}(\sigma_i, \sigma_j)\}$. By the same argument as in the proof of Lemma 3.B.6, there is a neighborhood U of p, a set $\{H(\sigma_i, \sigma_j)\}_{(\sigma_i,\sigma_j)\in S(p)}$ of hyperplanes in $\mathbb{R}^n$, and a homeomorphism $\phi : U \to \mathcal{B}$, where $\mathcal{B}$ is the open unit n-ball, such that

$$\phi\Big(\bigcap_{(\sigma_i,\sigma_j)\in S'(p)} I_{\boldsymbol{a}}(\sigma_i, \sigma_j) \cap U \Big) = \bigcap_{(\sigma_i,\sigma_j)\in S'(p)} H(\sigma_i, \sigma_j) \cap \mathcal{B}$$

for all $S'(p) \subseteq S(p)$. See Figure 3.B.1.

Because $M_{\boldsymbol{a},i} \pitchfork M_{\boldsymbol{a},j}$ for all $i \neq j$, we have that $\mathcal{B} \setminus H(\sigma_i, \sigma_j)$ is the disjoint union of open sets $W_1(\sigma_i, \sigma_j)$ and $W_2(\sigma_i, \sigma_j)$ such that

$$\begin{aligned} f_{\boldsymbol{a}}(\sigma_i, p') < f_{\boldsymbol{a}}(\sigma_j, p') \qquad & \text{for all } p' \in \phi^{-1}(W_1(\sigma_i, \sigma_j)) , \\ f_{\boldsymbol{a}}(\sigma_j, p') < f_{\boldsymbol{a}}(\sigma_i, p') \qquad & \text{for all } p' \in \phi^{-1}(W_2(\sigma_i, \sigma_j)) \end{aligned}$$

for all $(\sigma_i, \sigma_j) \in S(p)$. For each $(\sigma_i, \sigma_j) \in S(p)$, define the set $V(\sigma_i, \sigma_j)$ as in Equation 3.25, and define the set V as in Equation 3.26. The set $\phi^{-1}(V)$ is a nonempty subset of $U \cap Z_{\boldsymbol{a}}^{\prec_0}$. This implies that p is a limit point of $Z_{\boldsymbol{a}}^{\prec_0}$, so $Z_{\boldsymbol{a}}^{\prec_0} \subseteq \overline{Z_{\boldsymbol{a}}^{\prec_0}}$. Because $\prec$ is not the same as $\prec_0$, we have that $Z_{\boldsymbol{a}}^{\prec} \cap Z_{\boldsymbol{a}}^{\prec_0} = \emptyset$. Therefore, $Z_{\boldsymbol{a}}^{\prec_0} \subseteq \partial Z_{\boldsymbol{a}}^{\prec_0}$ and

$$\bigcup_{\prec \text{ in } \mathcal{O}} Z_{\boldsymbol{a}}^{\prec} \subseteq \partial Z_{\boldsymbol{a}}^{\prec_0} .$$

Now suppose that p is in the complement of $Z_{\boldsymbol{a}}^{\prec_0} \cup \left(\bigcup_{\prec \text{ in } \mathcal{O}} Z_{\boldsymbol{a}}^{\prec} \right)$. Because $\prec_{f_{\boldsymbol{a}}(\cdot,p)}$ is not the same as $\prec_0$ or any $\prec$ in $\mathcal{O}$, there is a pair (σ_i, σ_j) of simplices such that $f(\sigma_i, p) < f(\sigma_j, p)$ and either we have $\sigma_j \prec_0 \sigma_i$ or we have $\sigma_j \not\prec \sigma_i$ and $\sigma_i \not\prec \sigma_j$. By continuity of f, there is a neighborhood U_{σ_i,σ_j} of p such that $f_{\boldsymbol{a}}(\sigma_i, p') < f_{\boldsymbol{a}}(\sigma_j, p')$ for all $p' \in U_{\sigma_i,\sigma_j}$. Therefore, U_{σ_i,σ_j} is in the complement of $Z_{\boldsymbol{a}}^{\prec_0}$, so p is not in $\overline{Z_{\boldsymbol{a}}^{\prec_0}}$. This implies

$$\partial Z_{\boldsymbol{a}}^{\prec_0} \subseteq \bigcup_{\prec \text{ in } \mathcal{O}} Z_{\boldsymbol{a}}^{\prec} ,$$

which completes the proof. □

Acknowledgments

I am very grateful for discussions with Andrew Blumberg, which led to the investigation of monodromy in PD bundles. I also thank Ryan Grady and Karthik Viswanathan for helpful discussions.

CHAPTER 4

Computing Persistence Diagram Bundles

This chapter is adapted from [Hic22a].

4.1 Introduction

In Chapter 3, I introduced PD bundles, a framework for studying the PH of a set of filtrations that is parameterized by a topological space B. In ordinary persistent homology, one is given a filtration (e.g., a filtered complex), and one studies how the homology changes as the filtration parameter increases. For a PD bundle, we consider the case in which one is given not a single filtration, but rather a fibered filtration function, which is a set $\{f_p : \mathcal{K}^p \to \mathbb{R}\}_{p \in B}$ of filtration functions that is parameterized by some topological space B (the base space), where $\mathcal{K}^p$ is a simplicial complex for each $p \in B$. At each $p \in B$, the sublevel sets of f_p form a filtered complex. The associated PD bundle is the space of persistence diagrams $\mathrm{PD}(f_p)$ as they vary with $p \in B$.

4.1.1 Contributions

I generalize Cohen-Steiner et. al's algorithm for computing vineyards [CEM06] to an algorithm for efficiently computing PD bundles. I restrict to the case in which the PD bundle is *piecewise linear*. This means that B is a simplicial complex, $\mathcal{K}^p \equiv \mathcal{K}$ for all $p \in B$, and for every simplex $\sigma \in \mathcal{K}$, the function $f_\sigma(p) := f_p(\sigma)$ is linear in p on every simplex of B. The restriction to piecewise-linear PD bundles allows us to take advantage of results in compu-

tational geometry, such as the Bentley–Ottman planesweep algorithm [BCK08] for finding intersections of lines in a plane and algorithms for solving the point-location problem in a line arrangement [ST86]. An analogous piecewise-linear restriction was helpful for computing vineyards in [CEM06].

The idea of the algorithm is to subdivide the base space B into polyhedrons and compute a PD "template" for each polyhedron. The subdivision is given by Proposition 3.3.5 ([Hic22c]). For any $p \in B$, the persistence diagram $\mathrm{PD}(f_p)$ can then be computed in $\mathcal{O}(N)$ time from the PD template for the polyhedron that contains p, where N is the number of simplices in $\mathcal{K}$. By contrast, computing $\mathrm{PD}(f_p)$ from scratch takes $\mathcal{O}(N^3)$ time in the worst case.

The piecewise-linear restriction is reasonable for most applications. For example, suppose that we have a point cloud $X(t, \mu)$ whose coordinates depend on time t and a parameter $\mu \in \mathbb{R}$. If the data set arises from either real-world data collection or through numerical simulation, then we likely only know the coordinates of the point cloud $X(t, \mu)$ at a discrete set $\{t_i\}$ of time steps and a discrete set $\{\mu_j\}$ of system-parameter values. For every (t_i, μ_j), there is the filtration function $f_{(t_i,\mu_j)} : \mathcal{K} \to \mathbb{R}$ associated with the Vietoris–Rips filtration (or any other filtration) of $X(t_i, \mu_j)$. For the Vietoris–Rips filtration, $\mathcal{K}$ is the simplicial complex that contains a simplex for every subset of points in the point cloud. To obtain a fibered filtration function, we define B to be a triangulation of $[\min t_i, \max t_i] \times [\min \mu_j, \max \mu_j]$ whose vertices are $\{(t_i, \mu_j)\}_{ij}$. We can extend $\{f_{(t_i,\mu_j)}\}_{ij}$ to a fibered filtration function with base space B by defining the filtration values of a simplex σ via linear interpolation of $\{f_{(t_i,\mu_j)}(\sigma)\}_{ij}$. By construction, the resulting PD bundle is piecewise linear.

I give full details only for the case in which $\dim(B)$ is a triangulated surface, but I discuss the generalization to higher dimensions in Section 4.3.2. When the base space B is a triangulated surface, it is already an improvement over a vineyard because it allows three parameters in total: a filtration parameter as well as two parameters that locally parameterize B.

4.1.2 Related Work

PD bundles were introduced in [Hic22c] (Chapter 4) as a generalization of vineyards [CEM06]. The algorithm that I present in this chapter for computing PD bundles is a generalization of the algorithm in [CEM06]. In many ways, the algorithm in this chapter is also reminiscent of the Rivet algorithm for computing fibered barcodes of 2D multiparameter persistence modules [LW15].

4.2 Background

A PD bundle is the space of PDs that arises from a set of filtrations that are parameterized by some topological space B. (See Definitions 3.2.1 and 3.2.2 in Chapter 3.) In [CEM06], it was computationally easier to work with a piecewise-linear vineyard, which is a vineyard for a fibered filtration function of the form $f : \mathcal{K} \times [t_0, t_1] \to \mathbb{R}$ such that $f(\sigma, \cdot)$ is piecewise linear for all $\sigma \in \mathcal{K}$. (See the discussion at the end of Section 2.5.) In Chapter 3, we extended this to a definition of piecewise-linear PD bundles. (See Definition 3.2.4.) For the remainder of the chapter, we only consider piecewise-linear PD bundles.

For example, in Section 4.1 we considered a point cloud $X(t, \mu)$ whose coordinates depended on time $t \in \mathbb{R}$ and parameter $\mu \in \mathbb{R}$. Given only the coordinates of the point cloud at a discrete set $\{t_i\}$ of time steps and a discrete set $\{\mu_j\}$ of parameter values, we obtained a filtration function $f_{(t_i,\mu_j)}$ for every (t_i, μ_j). We extended this to a piecewise-linear fibered filtration function on $B = [\min t_i, \max t_i] \times [\min \mu_j, \max \mu_j]$ via linear interpolation of the filtration values for each simplex $\sigma \in \mathcal{K}$.

More generally, suppose that we are given a fibered filtration function of the form $f : \mathcal{K} \times \prod_{i=1}^{m} \mathcal{I}_i \to \mathbb{R}$, where each $\mathcal{I}_i$ is a finite subset of $\mathbb{R}$, and we wish to extend f to a fibered filtration function whose base is $B := \prod_{i=1}^{m} [\min \mathcal{I}_i, \max \mathcal{I}_i]$. First, we construct a triangulation B (i.e., an m-dimensional simplicial complex) of $\prod_{i=1}^{m} [\min \mathcal{I}_i, \max \mathcal{I}_i]$ whose set of vertices is $\prod_{i=1}^{m} \mathcal{I}_i$. (See [LRS10], for example, for a method to triangulate a cubical

complex.) We then extend f to a piecewise-linear fibered filtration function $f : \mathcal{K} \times B \to \mathbb{R}$ by linearly interpolating $f(\sigma, \cdot)$ on each simplex $\Delta \in B$ for all simplices $\sigma \in \mathcal{K}$.

In [Hic22c] (Chapter 3), it was shown that if $f : \mathcal{K} \times B \to \mathbb{R}$ is a piecewise-linear fibered filtration function on an n-dimensional simplicial complex B, then B can be subdivided into n-dimensional polyhedra such that within each polyhedron P, there is a "template" from which $PD_q(f(\cdot, p))$ can be computed for any $p \in P$. The template is a list of (birth, death) simplex pairs (σ_b, σ_d).

The polyhedra are defined as follows. We define

$$I(\sigma, \tau) := \{p \in B \mid f(\sigma, p) = f(\tau, p)\}.$$

For every n-simplex Δ in B, the intersection $I(\sigma, \tau) \cap \Delta$ is $\emptyset$, Δ, a vertex of Δ, or the intersection of an $(n-1)$-dimensional hyperplane with Δ. The set

$$\bigcup_{\Delta \in B} \partial\Delta \cup \left\{ \Big(I(\sigma, \tau) \cap \Delta \Big) \mid \emptyset \subset \Big(I(\sigma, \tau) \cap \Delta \Big) \subset \Delta \right\} \tag{4.1}$$

determines the boundaries of a polyhedral decomposition of B, where Δ denotes an n-simplex of B and $\partial\Delta$ denotes the boundary of Δ. Every polyhedron face is either a subset of $\partial\Delta$ or a subset of $I(\sigma, \tau) \cap \Delta$ for some simplex Δ of B.

As in Sections 2.4 and 2.5, we define the *simplex indexing induced by* f as follows. Let $\sigma_1, \ldots, \sigma_N$ be the simplices of $\mathcal{K}$, indexed such that $i < j$ if σ_i is a proper face of σ_j. We define the simplex indexing function $\mathrm{idx}_f : \mathcal{K} \times B \to \{1, \ldots, N\}$ to be the unique function such that $\mathrm{idx}_f(\sigma_i, p) < \mathrm{idx}_f(\sigma_j, p)$ if we have $f(\sigma_i, p) < f(\sigma_j, p)$ or we have $f(\sigma_i, p) = f(\sigma_j, p)$ and $i < j$.

Proposition 3.3.5 in Chapter 3 says that within each polyhedron P, the set $\{(\sigma_b, \sigma_d)\}$ of (birth, death) simplex pairs for $f(\cdot, p)$ is constant with respect to p.

4.3 Computing piecewise-linear PD bundles

The algorithm to compute piecewise-linear PD bundles has three main parts.

1. **Compute the polyhedra:** First, we compute the polyhedra on which the simplex indexing (and thus pairing function) is constant (see Proposition 3.3.5). For every pair P_1, P_2 of adjacent polyhedra, the face Q that they share is a subset of a set of the form $I(\sigma_{i_1}, \sigma_{j_1}) \cap \cdots \cap I(\sigma_{i_m}, \sigma_{j_m})$ for some m. In the "generic case," defined below at the beginning of Section 4.3.1.1, we have $m = 1$. We compute and store a reference that is associated with Q to the set $\{(\sigma_{i_1}, \sigma_{j_1}), \ldots, (\sigma_{i_m}, \sigma_{j_m})\}_{k=1}^{m}$.

2. **Compute the pairing function:** Within each polyhedron P, the set of (birth, death) simplex pairs for $f(\cdot, p)$ is constant with respect to $p \in P$. We compute the set of (birth, death) simplex pairs for each P. First, we choose an initial point $p_* \in B$ at which we compute the simplex indexing at p_*, the boundary matrix $D(p_*)$, an RU decomposition $D(p_*) = R(p_*)U(p_*)$, and the pairing function at p_*. (This is the pairing function for the entire polyhedron that contains p_*.) This takes $\mathcal{O}(N^3)$ time. We then traverse the polyhedra, starting with the polyhedron that contains p_* and visiting each polyhedron at least once. As we move from one polyhedron to the next via a shared face Q, we use the set $\{(\sigma_{i_1}, \sigma_{j_1}), \ldots, (\sigma_{i_m}, \sigma_{j_m})\}_{k=1}^{m}$ computed earlier for Q to update the RU decomposition and pairing function via the update rules that are used when computing vineyards (see [CEM06]). For each polyhedron, we store its pairing function (i.e., the pairs (σ_b, σ_d) of birth and death simplex pairs and also the unpaired simplices σ_b, which are birth simplices for homology classes that never die).

3. **Query the PD bundle:** To see the qth persistence diagram $\mathrm{PD}_q(f(\cdot, p))$ associated with point $p \in B$, we first locate the polyhedron P that contains p. For each pair (σ_b, σ_d) of simplices in the pairing function for P, the diagram $\mathrm{PD}_q(f(\cdot, p))$ has a point with coordinates $(f(\sigma_b, t), f(\sigma_d, t))$ if $\dim(\sigma_b) = q$. For every q-dimensional simplex σ_b that is unpaired in P, the diagram $\mathrm{PD}_q(f(\cdot, p))$ contains the point $(f(\sigma_b, t), \infty)$.

In what follows, we elaborate on each step of the algorithm above. We focus on the case in

which B is a triangulated surface.

4.3.1 Special case: B is a triangulated surface

Let $\mathcal{K}$ be a simplicial complex with simplices $\sigma_1, \ldots, \sigma_N$, indexed such that $i < j$ if σ_i is a proper face of σ_j. Let $f : \mathcal{K} \times B \to \mathbb{R}$ be a piecewise-linear fibered filtration function, and suppose that B is a triangulated surface. If T is a triangle in B, then $I(\sigma, \tau) \cap T$ is either $\emptyset$, T, a vertex of T, or a line segment whose endpoints are on ∂T. Figure 4.3.1 shows a few possible cases for $I(\sigma, \tau)$. The set in Equation 4.1 is a set L of line segments, and the polygonal subdivision induced by L is a *line arrangement*[1] $\mathcal{A}(L)$. For example, see Figure 3.1.

To simplify the exposition, we will make two genericity assumptions for the remainder of Section 4.3.1. The idea of the algorithm is not different in the general case, but it requires some technical modifications, which we discuss in Appendix 4.A. The assumptions are as follows:

1. For all distinct simplices $\sigma, \tau \in \mathcal{K}$ and all vertices $v \in B$, we have that $f(\sigma, v) \neq f(\tau, v)$. This implies that for all triangles $T \in B$, the intersection $I(\sigma, \tau) \cap T$ is either $\emptyset$ or a line segment whose endpoints are not vertices of T. For example, see Figures 4.3.1a and 4.3.1b.

2. For all distinct simplices $\sigma_{i_1}, \sigma_{j_1}, \sigma_{i_2}, \sigma_{j_2} \in \mathcal{K}$ and every triangle $T \in B$ such that $I(\sigma_{i_1}, \sigma_{j_1}) \cap T$ and $I(\sigma_{i_2}, \sigma_{j_2}) \cap T$ are nonempty, the line segments $I(\sigma_{i_1}, \sigma_{j_1}) \cap T$ and $I(\sigma_{i_2}, \sigma_{j_2}) \cap T$ do not share any endpoints.

Throughout Section 4.3.1, we use the following notation. Let m denote the number of triangles in B. The numbers of vertices and edges in B are both $\mathcal{O}(m)$. Let N denote

[1] Usually, a "line arrangement" refers to a planar subdivision that is induced by a set of lines in the plane. However, here I am using the term "line arrangement" slightly more generally. Every line segment in L lies in some triangle T of B, and the line segments subdivide each triangle into polygons.

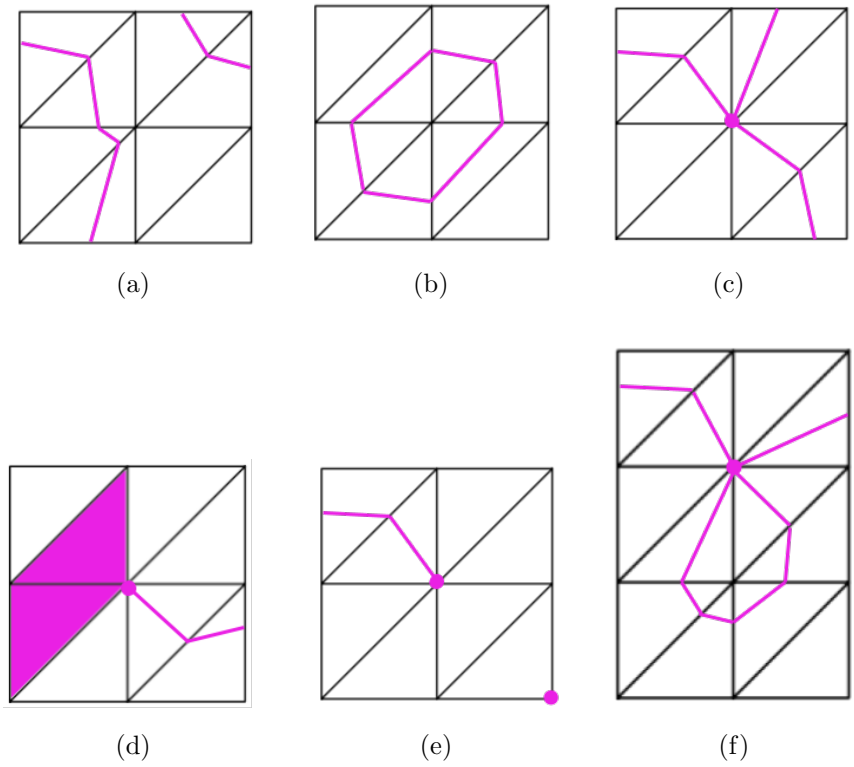


Figure 4.3.1: A few possible cases for the set $I(\sigma, \tau)$, which are shown in pink. The black lines are the 1-skeleton of B.

the number of simplices in $\mathcal{K}$. For every triangle T of B, let κ_T denote the number of line segments ℓ of the form $I(\sigma, \tau) \cap T$ and let $\kappa := \sum_{T \in B} \kappa_T$. The worst case is $\kappa_T = \mathcal{O}(N^2)$ and $\kappa = \mathcal{O}(mN^2)$, but these are very crude upper bounds. Let μ_T denote the number of vertices of $\mathcal{A}(L)$ in the interior of triangle $T \in B$; the quantity μ_T is equal to the number of points of the form $I(\sigma_{i_1}, \sigma_{j_1}) \cap I(\sigma_{i_2}, \sigma_{j_2}) \cap T$. Let μ denote the total number of vertices in $\mathcal{A}(L)$, which is $\sum_{T \in B} \mu_T + \mathcal{O}(m + \kappa)$. In the worst case, $\mu_T = \mathcal{O}(\kappa_T^2) = \mathcal{O}(N^4)$ and $\mu = \mathcal{O}(mN^4)$, but these are again very crude upper bounds. The numbers of edges and polygons in $\mathcal{A}(L)$ are $\mathcal{O}(\mu)$.[2]

4.3.1.1 Computing the polygons

For a piecewise-linear vineyard, computing the intervals on which the simplex indexing is constant can be reduced to finding the intersections between the piecewise-linear functions $y = f(\sigma, p)$ and $y = f(\tau, p)$ for all pairs (σ, τ) of simplices in $\mathcal{K}$. Likewise, for a piecewise-linear PD bundle, computing the polygons on which the simplex indexing is constant can be reduced to finding the intersections $I(\sigma, \tau)$ for all pairs (σ, τ) of simplices.

We seek to compute the line arrangement $\mathcal{A}(L)$, where L is the set of line segments defined by Equation 4.1. (See Figure 3.1.) The polygons of $\mathcal{A}(L)$ are the polygons on which the simplex indexing is constant. We store $\mathcal{A}(L)$ using a doubly-connected edge list (DCEL) data structure, which is a standard data structure for storing a polygonal subdivision of the plane [BCK08]. The DCEL data structure can be used without modification to represent $\mathcal{A}(L)$, which is a polygonal subdivision of a triangulated surface.[3] The space complexity of $\mathcal{A}(L)$ is $\mathcal{O}(\mu)$. We compute $\mathcal{A}(L)$ using the following algorithm (illustrated in Figure 4.3.2):

[2]Within each triangle T of B, the numbers of edges and polygons in $\mathcal{A}(L) \cap T$ are $\mathcal{O}(\mu_T)$ by Euler's formula, as noted in [BCK08].

[3]The primary reason to consider triangulated surfaces, rather than any simplicial complex B such that $\dim(B) \leq 2$, is so that we can use a DCEL data structure to represent $\mathcal{A}(L)$. Otherwise, what follows in Section 4.3.1 works just as well for any 2D simplicial complex.

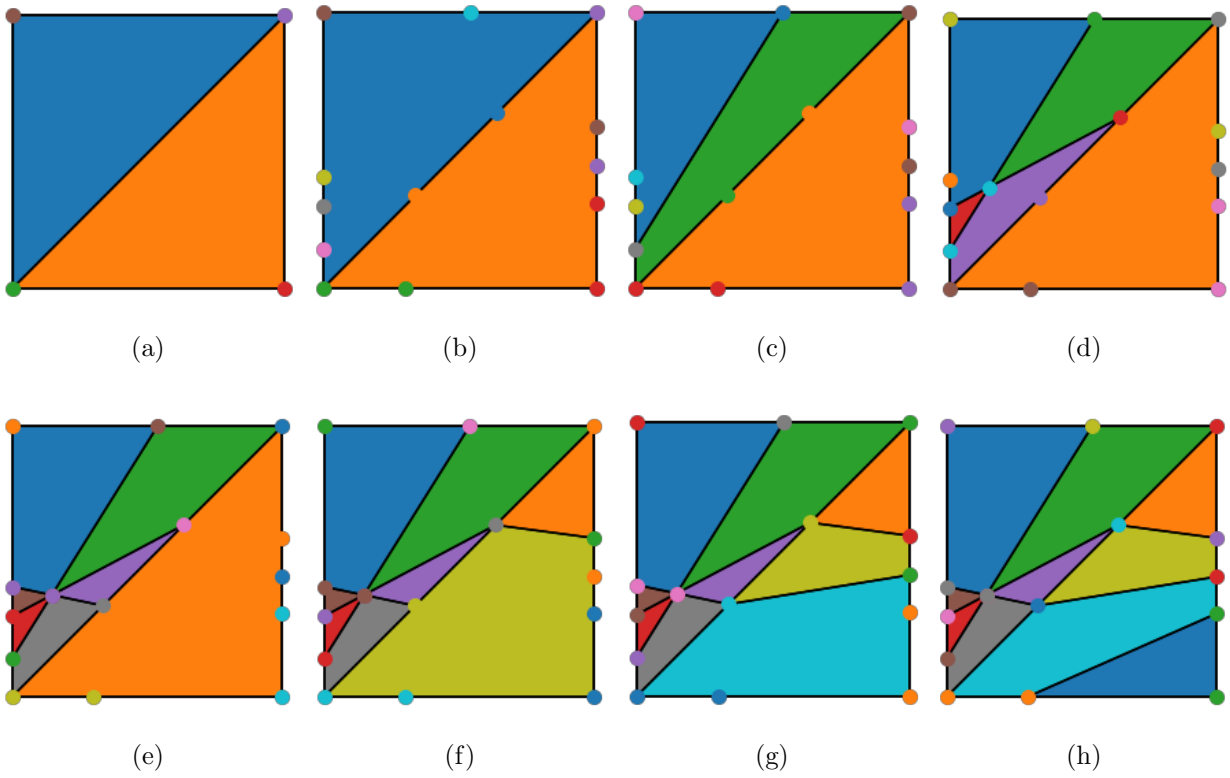


Figure 4.3.2: Computing the polygons. (A) The line arrangement $\mathcal{A}(L)$ is initialized to represent the triangulated base space B, which in this case consists of two triangles. (B) We find the vertices v of $\mathcal{A}(L)$ that lie on the 1-skeleton of B. (C)–(H) For each triangle T of B, we incrementally add the line segments of the form $I(\sigma, \tau) \cap T$. The endpoints of a given line segment are a pair (v, w) of vertices in (B). In (D), an internal vertex (a vertex at the intersection of two line segments) is created when the last line segment is added. Three of the line segments intersect at the internal vertex.

1. We initialize $\mathcal{A}(L)$ so that it represents the triangulation B. (See Figure 4.3.2a.) In addition to the usual data that a DCEL stores, we enumerate the triangles in B and every half edge e stores the index of the triangle in B that e is on the boundary of.

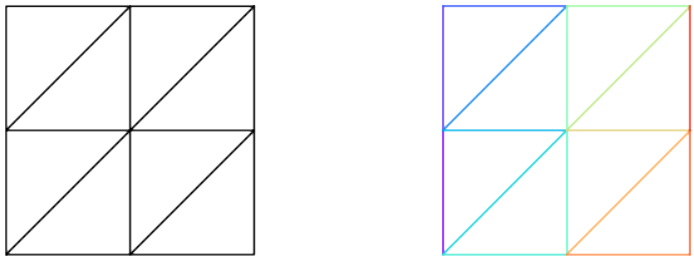
2. For every triangle $T \in B$, we initialize an empty dictionary $\mathcal{D}(T)$.[4] The keys will be pairs (σ, τ) of simplices for which $I(\sigma, \tau) \cap T$ is a line segment, and the value of (σ, τ) will be a list of the endpoints (v, w) of the line segment. We denote the value of (σ, τ) by $\mathcal{D}(T)[(\sigma, \tau)]$. These dictionaries use $\mathcal{O}(\kappa)$ space.

3. For each edge e of B, we compute the vertices of $\mathcal{A}(L)$ that lie on e. (See Figure 4.3.2b.) These are the vertices that equal $I(\sigma, \tau) \cap e$ for some pair (σ, τ) of simplices. To do this, we consider the restriction of f to e, which is a 1-parameter filtration function (the input to a vineyard). For each $\sigma \in \mathcal{K}$, the set $\{(p, f(\sigma, p)) \mid p \in e\}$ is a line segment $\ell_{\sigma,e}$ and $I(\sigma, \tau) \cap e$ is the point $p \in e$ at which $\ell_{\sigma,e}$ and $\ell_{\tau,e}$ intersect. We use the Bentley–Ottman planesweep algorithm [BCK08] to compute these intersections, thus obtaining the vertices v of $\mathcal{A}(L)$ that lie on e. For a vertex v that equals $I(\sigma, \tau) \cap e$, we add v to the list $\mathcal{D}(T)[(\sigma, \tau)]$. Completing step 3 takes $\mathcal{O}(N)$ space and $\mathcal{O}((mN + \kappa) \log N)$ time in total for all edges in B, and can be parallelized over the edges.[5]

4. For each triangle $T \in B$ and for each key (σ, τ) in the dictionary $\mathcal{D}(T)$, there is an associated pair (v, w) of vertices which are the endpoints of the line segment $I(\sigma, \tau) \cap T$. We seek to add all of these line segments to the DCEL that represents $\mathcal{A}(L)$. (See Figures 4.3.2C–H.) There are many standard algorithms for doing this: one example is the incremental algorithm (see e.g., Chapter 8.3 of [BCK08]), in which the line segments are incrementally added one at a time. The worst-case run time of the incremental

[4] A dictionary is a data structure for storing (key, value) pairs.

[5] In some cases, in may be more efficient to consider the restriction of f to an Euler path γ through the 1-skeleton of B, rather than the restriction of f to each edge separately. For example, if B is of the form in Figure 4.3.3a, then an Euler path is given by Figure 4.3.3b.



(a) (b)

Figure 4.3.3: (A) A triangulated base space B. (B) An Euler path γ through the 1-skeleton of B, starting at the bottom-left vertical edge (violet) and ending at the top-right vertical edge (red).

algorithm in triangle T is $\mathcal{O}(\kappa_T^2)$, for a total run time of $\mathcal{O}(\sum_{T \in B} \kappa_T^2)$, and it can be parallelized over the triangles $T \in B$.

In Figure 4.3.2, we illustrate the algorithm for computing the polygons.

As we add line segments to $\mathcal{A}(L)$, we keep track of the pairs (σ, τ) of simplices that correspond to each edge of $\mathcal{A}(L)$. If edge e of $\mathcal{A}(L)$ is a subset of $I(\sigma, \tau)$, then e stores a reference to the pair (σ, τ). We add the reference to (σ, τ) at the time that edge e is created in $\mathcal{A}(L)$. If P_1, P_2 are adjacent polygons of $\mathcal{A}(L)$ that share edge e, then the simplex indexings in P_1, P_2 are related via the transposition of σ and τ.

4.3.1.2 Computing the pairing function

Let G be the dual graph to the line arrangement $\mathcal{A}(L)$. The graph G contains a vertex v_P for every polygon P of $\mathcal{A}(L)$ and an edge between two vertices if the corresponding polygons are adjacent. We compute a path Γ that visits every vertex of G at least once. For example, see Figure 4.3.4. One way to obtain such a path is via depth-first search, which takes $\mathcal{O}(\mu)$ time because the number of nodes in G (polygons of $\mathcal{A}(L)$) is $\mathcal{O}(\mu)$. This yields a path Γ whose length is $\mathcal{O}(\mu)$.

Figure 4.3.4: A path Γ (in white) that visits every polygon in the line arrangement $\mathcal{A}(L)$.

At the first vertex v_P of Γ, we compute the simplex indexing in polygon P, the RU decomposition for the boundary matrix in P, and the (birth, death) simplex pairs in P. To store the RU decomposition, we use the sparse matrix data structure from [CEM06]. The polygon P stores a reference to its (birth, death) simplex pairs. To store the current simplex indexing, each simplex stores a reference to its index in the current indexing (which we initialize to the indexing in P).

We traverse the path Γ. As we walk from one polygon P_1 to the next polygon P_2 by crossing an edge e in $\mathcal{A}(L)$, we update the simplex indexing, the RU decomposition, and the (birth, death) simplex pairs. To update the simplex indexing, recall that edge e stores a reference to the simplex pair (σ, τ) such that $e \subseteq I(\sigma, \tau)$. This implies that the simplex indexings in P_1 and P_2 are related via the transposition of σ and τ because we must have (without loss of generality) $f(\sigma, p) > f(\tau, p)$ for $p \in P_1$ and $f(\sigma, p) < f(\tau, p)$ for $p \in P_2$, with $f(\sigma, p) = f(\tau, p)$ on the shared edge e. We update the simplex indexing by swapping the indices that σ and τ store. To update the RU decomposition and the (birth, death) simplex pairs, we apply the update algorithm of [CEM06]. This takes $\mathcal{O}(N)$ time in the worse case, but often approximately constant time in practice; see [CEM06] and the earlier discussion in Section 2.5. In P_2, we store the new (birth, death) simplex pairs. Storing the simplex pairs uses $\mathcal{O}(N)$ space for each polygon of $\mathcal{A}(L)$, so we use $\mathcal{O}(N\mu)$ space in total.

We can optimize the space requirements by recalling from [CEM06] that most updates of the simplex indexing do not change the (birth, death) simplex pairs. If the update from P_1 to P_2 does not change the simplex pairs, we can delete the edge in $\mathcal{A}(L)$ that P_1 and P_2 share, thus merging P_1 and P_2 into a single polygon and reducing the size of $\mathcal{A}(L)$.

4.3.1.3 Querying the PD bundle

We consider the scenario in which a user seeks to query many points $p \in B$ in real time and see the qth persistence diagram $\mathrm{PD}_q(f(\cdot, p))$ associated with each queried point p.

To compute the qth persistence diagram $\mathrm{PD}_q(f(\cdot, p))$ associated with a given p, we first identify the polygon P of $\mathcal{A}(L)$ that contains p. We do this in two steps. The first step is to identify the triangle $T \in B$ that contains p. This takes $\mathcal{O}(m)$ worst-case time because it takes constant time to test if a given triangle contains p. In certain cases, one can identify the triangle T more efficiently. For example, if B is a triangulation of the form in Figure 4.3.3a, then one can locate the triangle T in $\mathcal{O}(1)$ time by simply examining the coordinates of p. The second step is to locate the polygon P in T that contains p. This is a well-studied problem in computational geometry; it is known as the *point-location problem*. When one is planning to perform many point-location queries on the same line arrangement (i.e., if one is querying many points $p \in B$), the standard strategy is to precompute a data structure so that the subsequent point-location queries can be done efficiently. There are many strategies for doing this (see, e.g., Chapter 38 in [TOG17]). One method is the slab-and-persistence method [ST86], in which one precomputes a "persistent search tree" for the line arrangement. We construct a persistent search tree for each triangle in B. Using separate persistent search trees for the planar subdivisions in each triangle, the slab and persistence method takes $\mathcal{O}(\sum_{T \in B} \mu_T \log(\mu_T))$ preprocessing time, $\mathcal{O}(\mu)$ space, and $\mathcal{O}(\max_{T \in B} \log \mu_T)$ time per query.

We obtain $\mathrm{PD}_q(f(\cdot, p))$ by evaluating $f(\cdot, p)$ on the simplices in the pairing function for polygon P, which was precomputed in the previous step (see Section 4.3.1.2). This takes $\mathcal{O}(N)$ time. For every (birth, death) pair (σ_b, σ_d) of simplices, $\mathrm{PD}_q(f(\cdot, p))$ has a point with

coordinates $(f(\sigma_b, t), f(\sigma_d, t))$ if $\dim(\sigma_b) = q$. For every unpaired q-dimensional simplex σ_b, the diagram $\mathrm{PD}_q(f(\cdot, p))$ has a point with coordinates $(f(\sigma_b, t), \infty)$.

4.3.2 Generalizing to higher-dimensional base spaces

The algorithm of Section 4.3.1 (as outlined at the beginning of Section 4.3) does not require many modifications for higher-dimensional base spaces B. We replace the subdivision of B into polygons by a subdivision of B into n-dimensional polyhedra, where $n = \dim(B)$. In the third step (querying the PD bundle), one uses a point-location algorithm for a hyperplane arrangement (see, e.g., [Mei93,CF94]). Only the first step (computing the polyhedra) requires a meaningful modification, which are described below.

When $n = 2$, the intersection of $I(\sigma, \tau)$ with a triangle $T \in B$ is the intersection of a line with T, which is a line segment $L_{\sigma,\tau,T}$. These line segments completely determine the polygonal subdivision of B because the line segments are the faces of the polygons. In turn, each line segment $L_{\sigma,\tau,T}$ is completely determined by the intersection of $L_{\sigma,\tau,T}$ with the 1-skeleton of B; this intersection is a pair $(v_{\sigma,\tau,T}, w_{\sigma,\tau,T})$ of points. In Section 4.3.1.1, we computed the set $\{(v_{\sigma,\tau,T}, w_{\sigma,\tau,T})\}_{\sigma,\tau,T}$ by restricting the fibered filtration function f to each edge of B and applying the Bentley–Ottman planesweep algorithm.

In general, the intersection of $I(\sigma, \tau)$ with an n-simplex $\Delta \in B$ is the intersection of an $(n-1)$-dimensional hyperplane $H_{\sigma,\tau,\Delta}$ with Δ; the intersection is an $(n-1)$-dimensional polyhedron $P_{\sigma,\tau,\Delta}$. The set $\{P_{\sigma,\tau,\Delta}\}_{\sigma,\tau,\Delta}$ completely determines the polyhedral subdivision of B that is given by Proposition 3.3.5 because the polyhedra $P_{\sigma,\tau,\Delta}$ are the $(n-1)$-dimensional faces of the n-dimensional polyhedra in the subdivision. In turn, each polyhedron $P_{\sigma,\tau,\Delta}$ is completely determined by its intersection with the edges of B, as follows. The m-dimensional faces of $P_{\sigma,\tau,\Delta}$ are the set

$$\{H_{\sigma,\tau,\Delta} \cap \Delta^{(m+1)} \mid \Delta^{(m+1)} \text{ is an } (m+1)\text{-dimensional face of } \Delta \text{ and } H_{\sigma,\tau,\Delta} \cap \Delta^{(m+1)} \neq \emptyset\}.$$

For every $(m+1)$-dimensional face $\Delta^{(m+1)}$ of Δ such that $H_{\sigma,\tau,\Delta} \cap \Delta^{(m+1)} \neq \emptyset$, the intersection

$H_{\sigma,\tau,\Delta} \cap \Delta^{(m+1)}$ is the m-dimensional polyhedron whose $(m-1)$-dimensional faces are the set

$$\{H_{\sigma,\tau,\Delta} \cap \Delta^{(m)} \mid \Delta^{(m)} \text{ is an } m\text{-dimensional face of } \Delta^{(m+1)}\}\,.$$

By induction, the faces of $P_{\sigma,\tau,\Delta}$ are determined by

$$\{H_{\sigma,\tau,\Delta} \cap e \mid e \text{ is a 1-dimensional face of } \Delta \text{ (i.e., } e \text{ is an edge)}\}\,,$$

which are the vertices of $P_{\sigma,\tau,\Delta}$. Consequently, we can compute each $P_{\sigma,\tau,\Delta}$ by determining its vertices. As in the case in which B is a triangulated surface, we do this by restricting the fibered filtration function f to each edge of B and applying the Bentley–Ottman planesweep algorithm.

4.4 Conclusions and Discussion

I introduced an algorithm for efficiently computing persistence diagram (PD) bundles when the fibered filtration function is piecewise linear. I gave full details for the case in which the base space B is a triangulated surface. Additionally, in Section 4.3.2, I discussed how one can generalize the algorithm to higher dimensions. I conclude with some questions and proposals for future work:

- What invariants can we use to summarize and analyze PD bundles in ways that do not require exploratory data analysis?

  The current algorithm asks a user to "query" the PD bundle at various points in the base space. This is useful for qualitative analysis or if one has a function whose input is a PD and one seeks to optimize that function over B. However, other applications may require global invariants.

- How do we generalize the algorithm to fibered filtration functions that are not piecewise linear?

For piecewise-linear fibered filtration functions, we used the fact that the base space B can be subdivided into polyhedrons such that there is a single PD "template" (a list of (birth, death) simplex pairs) for each polyhedron. The template can then be used to obtain $\mathrm{PD}_q(f_p)$ at any point p in the polyhedron. For "generic" fibered filtration functions, it was shown in [Hic22c] (Chapter 3) that the base space B is stratified such that for each stratum, there is a single PD template that can be used to obtain $\mathrm{PD}_q(f_p)$ at any point p in the stratum.

- How do we generalize to the case where $\mathcal{K}^p$ is not constant with respect to $p \in B$?

  In this case, simplices are added and removed from the filtration as $p \in B$ varies, so the algorithm must be modified.

APPENDIX

4.A Technical details of the algorithm

Let $\mathcal{K}$ be a simplicial complex with N simplices, indexed $\sigma_1, \ldots, \sigma_N$ such that $i < j$ if σ_i is a proper face of σ_j. Let B be a triangulated surface, and let $f : \mathcal{K} \times B \to \mathbb{R}$ be a piecewise-linear fibered filtration function. In Section 4.3.1, we made two generic assumptions to simplify the exposition. If assumption (1) holds, then every nonempty $I(\sigma_i, \sigma_j) \cap T$ is a line segment ℓ that subdivides triangle T into polygons Q_1, Q_2 such that

$$(\mathrm{idx}_f(\sigma_i, p_1) - \mathrm{idx}_f(\sigma_j, p_1)) \times (\mathrm{idx}_f(\sigma_i, p_2) - \mathrm{idx}_f(\sigma_j, p_2)) < 0$$

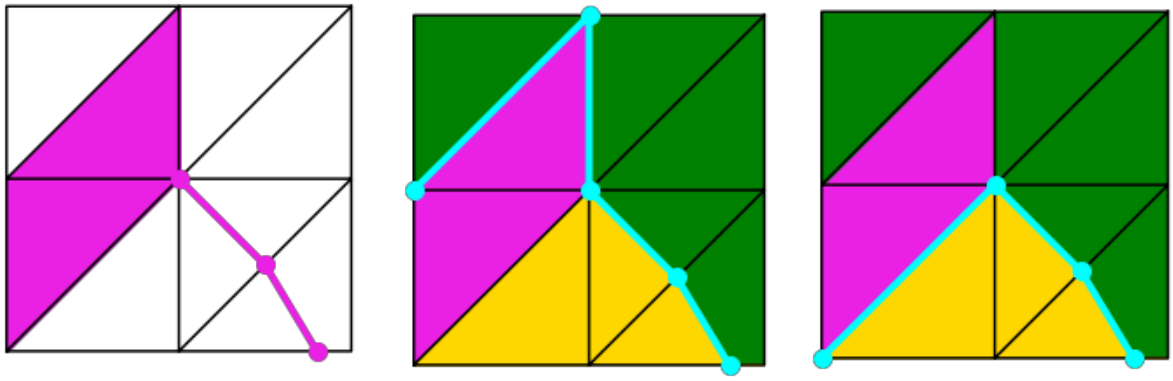
for all $p_1 \in Q_1$ and $p_2 \in Q_2$ (i.e., σ_i and σ_j have different relative orders in Q_1 and Q_2). We say that σ_i and σ_j *swap along* ℓ because σ_i and σ_j have different relative orders on either side of ℓ. If assumption (1) does not hold, then for any triangle T in B and pair (σ_i, σ_j) of simplices, it is possible that $I(\sigma_i, \sigma_j) \cap T$ equals either T or an edge of T. If e is an edge of triangle T such that $e \subseteq I(\sigma_i, \sigma_j) \cap T$, then σ_i and σ_j *swap along line segment* e if

$$(\mathrm{idx}_f(\sigma_i, p_1) - \mathrm{idx}_f(\sigma_j, p_1)) \times (\mathrm{idx}_f(\sigma_i, p_2) - \mathrm{idx}_f(\sigma_j, p_2)) < 0$$

for all $p_1 \in T_1$ and $p_2 \in T_2$, where T_1 and T_2 are the triangles of B that are adjacent to e (i.e., σ_i and σ_j have different relative orders in T_1 and T_2). In Figure 4.A.1, we illustrate an example where assumption (1) does not hold. We highlight the line segments that σ_i and σ_j swap along.

If assumption (2) does not hold, then there may be a line segment ℓ in a triangle T such that $I(\sigma_{i_1}, \sigma_{j_1}) \cap T = \ell = I(\sigma_{i_2}, \sigma_{j_2}) \cap T$ for two distinct pairs $(\sigma_{i_1}, \sigma_{j_1})$ and $(\sigma_{i_2}, \sigma_{j_2})$ of simplices in $\mathcal{K}$.

In Sections 4.A.2 and 4.A.3, I explain the modifications for the algorithm when we do not make the assumptions of Section 4.3.1. It suffices to modify step 1 (see Section 4.3.1.1) and step 2 (see Section 4.3.1.2). Step 3 (Section 4.3.1.3) remains the same.



(a) (b) (c)

Figure 4.A.1: Examples of fibered filtration functions for which assumption (1) of Section 4.3.1 does not hold. (A) The pair (σ_i, σ_j) is a pair of simplices such that $I(\sigma_i, \sigma_j) \cap T = T$ for every pink triangle T and $I(\sigma_i, \sigma_j) \cap T = \ell$ if $\ell \subseteq T$ is a pink line segment. Without loss of generality $i < j$, so $\mathrm{idx}_f(\sigma_i, p) < \mathrm{idx}_f(\sigma_j, p)$ if $f(\sigma_i, p) = f(\sigma_j, p)$. (B) Suppose that $f(\sigma_j, p) < f(\sigma_i, p)$ on green triangles and $f(\sigma_i, p) < f(\sigma_j, p)$ on yellow triangles. In blue, we draw the line segments on which the pair (σ_i, σ_j) swaps. (C) Suppose that $f(\sigma_i, p) < f(\sigma_j, p)$ on green triangles and $f(\sigma_j, p) < f(\sigma_i, p)$ on yellow triangles. In blue, we again draw the line segments on which the pair (σ_i, σ_j) swaps.

4.A.1 Preliminaries

As in Section 4.3.1.1, we consider the restriction of f to every edge e of B to find the vertices of $\mathcal{A}(L)$ that lie on the 1-skeleton of B.

Definition 4.A.1. A vertex v for a pair (σ, τ) of simplices is *detected along edge* e if, while traversing edge e during the Bentley-Ottman algorithm, we detect the point $v \in e$ as a point where the relative order of σ and τ changes.

A vertex v for the pair (σ, τ) is detected along edge e if and only if σ and τ have different relative orders at the endpoints of e. If v is an endpoint of e, then v is detected if and only if the relative order at v is different from the relative order in the interior of the edge e.

Definition 4.A.2. A line segment (v, w) is *detected in triangle* T if there is a pair (σ, τ) of simplices such that vertex v is detected along an edge e_1 of T for (σ, τ) and vertex w is detected along an edge e_2 of T for (σ, τ).

Lemma 4.A.3, below, characterizes the conditions under which a pair of simplices swaps along a line segment.

Lemma 4.A.3. Let (v, w) be a line segment that is not on the boundary of B.

1. If v and w are not the endpoints of an edge in B, let T be the unique triangle that contains (v, w). A pair (σ_i, σ_j) of simplices swaps along the line segment (v, w) if and only if (v, w) is detected in triangle T.

2. If v and w are the endpoints of an edge in B, let T_1, T_2 be the two triangles adjacent to that edge. A pair (σ_i, σ_j) of simplices swaps along the line segment (v, w) if and only if (v, w) is detected in exactly one of T_1, T_2.

Proof. Statement (1) was the situation in Section 4.3.1, so it remains only to prove statement (2).

Suppose that (v, w) are the endpoints of an edge e in B that is not on the boundary of B. Let T_1, T_2 be the two triangles adjacent to T. As illustrated in Figure 4.A.2a, we denote the third vertex of T_1 by u_1, the third vertex of T_2 by u_2, the other two edges in T_1 by e_2, e_3, and the other two edges in T_2 by e_4, e_5.

A pair (σ_i, σ_j) swaps along (v, w) only if $e \subseteq I(\sigma_i, \sigma_j) \cap T_1 \cap T_2$. If $e \subseteq I(\sigma_i, \sigma_j) \cap T$ for triangle T, then either $I(\sigma_i, \sigma_j) \cap T = e$ or $I(\sigma_i, \sigma_j) \cap T = T$. If we have both $I(\sigma_i, \sigma_j) \cap T_1 = T_1$ and $I(\sigma_i, \sigma_j) \cap T_2 = T_2$, then (σ_i, σ_j) does not swap along (v, w) because σ_i and σ_j have the same relative order in T_1 and T_2. Therefore, the pair (σ_i, σ_j) swaps along (v, w) only if the intersection of $I(\sigma_i, \sigma_j)$ with one triangle is e and the intersection with the other triangle is either e or the entire triangle. Without loss of generality, $I(\sigma_i, \sigma_j) \cap T_1 = e$ and either $I(\sigma_i, \sigma_j) \cap T_2 = T_2$ or $I(\sigma_i, \sigma_j) \cap T_2 = e$.

The line segment (v, w) can only be detected in triangle T_k if $e = I(\sigma_i, \sigma_j) \cap T_k$. If $I(\sigma_i, \sigma_j) \cap T_1 = e$ then we must also have $e \subseteq I(\sigma_i, \sigma_j) \cap T_2$, so either $I(\sigma_i, \sigma_j) \cap T_2 = T_2$ or $I(\sigma_i, \sigma_j) \cap T_2 = e$ (and vice versa if $I(\sigma_i, \sigma_j) \cap T_2 = e$). Therefore (v, w) is detected in T_k only if $I(\sigma_i, \sigma_j) = e$ and the intersection with the other triangle is either e or the whole triangle. Without loss of generality, $T_k = T_1$.

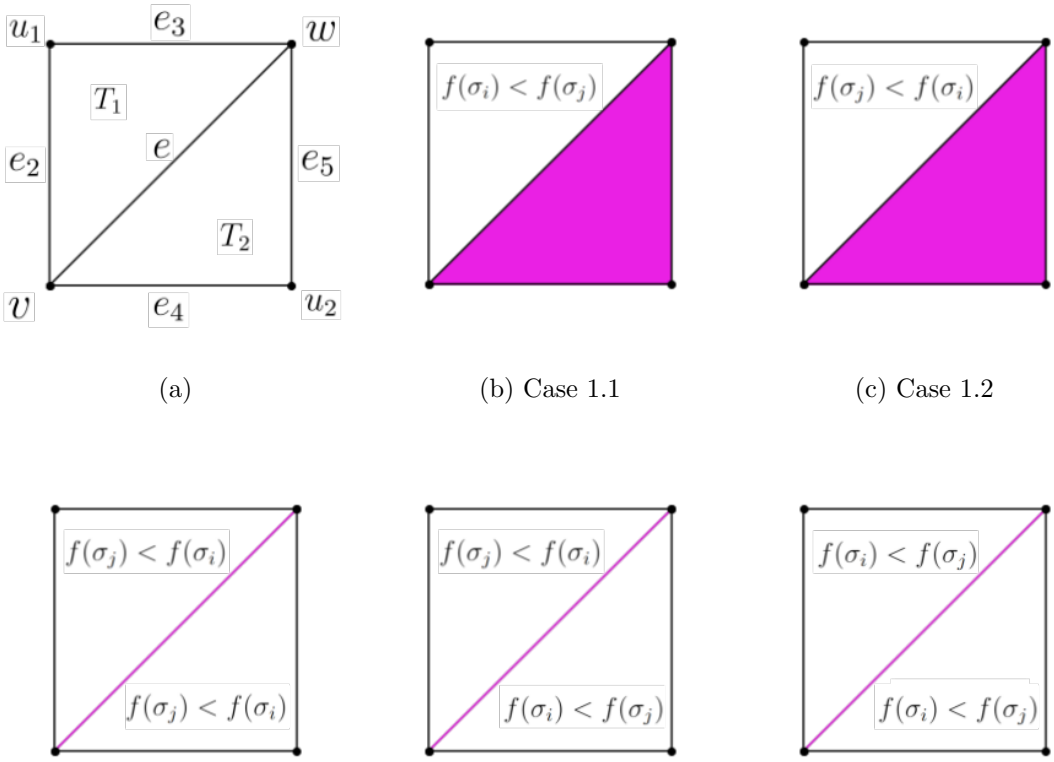
In Figures 4.A.2b–4.A.2f, we illustrate the possible cases in which $I(\sigma_i, \sigma_j) \cap T_1 = e$ and either $I(\sigma_i, \sigma_j) \cap T_2 = T_2$ or $I(\sigma_i, \sigma_j) \cap T_2 = e$. We will show that in each of these cases, statement (2) holds. In all other cases, we have already shown that neither (σ_i, σ_j) swaps along the line segment (v, w) nor is (v, w) detected in T_1 or T_2.

Without loss of generality, we assume $i < j$, so $\mathrm{idx}_f(\sigma_i, p) < \mathrm{idx}_f(\sigma_j, p)$ if $p \in I(\sigma_i, \sigma_j)$.

Case 1: $I(\sigma_i, \sigma_j) \cap T_2 = T_2$.

There are two subcases.

1. **Case 1.1:** (Figure 4.A.2b) $f(\sigma_i, p) < f(\sigma_j, p)$ for all $p \in T_1 \setminus e$.



(a) (b) Case 1.1 (c) Case 1.2

(d) Case 2.1 (e) Case 2.2 (f) Case 2.3

Figure 4.A.2: (A) The vertices, edges, and triangles that were defined in the proof of Lemma 4.A.3. (B–F) The cases in the proof of Lemma 4.A.3. Pink regions are regions on which σ_i and σ_j have equal filtration values.

In this subcase, we have $\text{idx}_f(\sigma_i, p) < \text{idx}_f(\sigma_j, p)$ for all $p \in T_1 \cup T_2$. Therefore, the pair (σ_i, σ_j) does not swap along (v, w). Neither v nor w is detected along any edges of T_1 or T_2, so the line segment (v, w) is not detected in either T_1 or T_2.

2. **Case 1.2:** (Figure 4.A.2c) $f(\sigma_j, p) < f(\sigma_i, p)$ for all $p \in T_1 \setminus e$.

   In this subcase, we have $\text{idx}_f(\sigma_i, p) < \text{idx}_f(\sigma_j, p)$ for $p \in T_2$ and $\text{idx}_f(\sigma_j, p) < \text{idx}_f(\sigma_i, p)$ for all $p \in T_1 \setminus e$. Therefore, the pair (σ_i, σ_j) swaps along (v, w). The vertex v is detected along edge e_2 and the vertex w is detected along the edge e_3. Because e_2 and e_3 are edges of T_1, the line segment (v, w) is detected in T_1. The vertices v and w are not detected along any edge of T_2, so (v, w) is not detected in T_2.

Case 2: $I(\sigma_i, \sigma_j) \cap T_2 = e$.

There are three subcases.

1. **Case 2.1:** (Figure 4.A.2d) $f(\sigma_j, p) < f(\sigma_i, p)$ for all $p \in T_1 \cup T_2 \setminus e$.

   In this subcase, we have $\text{idx}_f(\sigma_j, p) < \text{idx}_f(\sigma_i, p)$ for all $p \in (T_1 \cup T_2) \setminus e$. Therefore, the pair (σ_i, σ_j) does not swap along (v, w). The vertex w is detected along edges e_3 and e_5. The vertex v is detected along edges e_2 and e_4. Therefore, (v, w) is detected in both T_1 and T_2.

2. **Case 2.2:** (Figure 4.A.2e) Either we have $f(\sigma_j, p) < f(\sigma_i, p)$ for all $p \in T_1 \setminus e$ and $f(\sigma_i, p) < f(\sigma_j, p)$ for all $p \in T_2 \setminus e$ or we have $f(\sigma_i, p) < f(\sigma_j, p)$ for all $p \in T_1 \setminus e$ and $f(\sigma_j, p) < f(\sigma_i, p)$ for all $p \in T_1 \setminus e$. Without loss of generality, we assume the former.

   In this subcase, $\text{idx}_f(\sigma_i, p) < \text{idx}_f(\sigma_j, p)$ for all $p \in T_2$ and $\text{idx}_f(\sigma_j, p) < \text{idx}_f(\sigma_i, p)$ for all $p \in T_1 \setminus e$. Therefore, the pair (σ_i, σ_j) swaps along (v, w). The vertex v is detected along e_2 and the vertex w is detected along e_3, so (v, w) is detected in triangle T_1. Neither v nor w is detected along any edge of T_2, so (v, w) is not detected in T_2.

3. **Case 2.3:** (Figure 4.A.2f) $f(\sigma_i, p) < f(\sigma_j, p)$ for all $p \in T_1 \cup T_2 \setminus e$.

   In this subcase, we have $\mathrm{idx}_f(\sigma_i, p) < \mathrm{idx}_f(\sigma_j, p)$ for all $p \in T_1 \cup T_2$. Therefore, the pair (σ_i, σ_j) does not swap along (v, w). Neither v nor w is detected along any edge of T_1 or T_2, so (v, w) is not detected in either T_1 or T_2.

□

Lemma 4.A.4 below will be used to modify step 2 of the algorithm: computing the simplex pairing function.

Lemma 4.A.4. Let $\mathrm{idx}_0, \mathrm{idx}_1 : \mathcal{K} \to \{1, \ldots, N\}$ denote two different simplex indexings (not necessarily compatible with f; see Definition 2.4.2), where N is the number of simplices in $\mathcal{K}$. Let $\{(\sigma_{i_k}, \sigma_{j_k})\}_{k=1}^m$ be the set of pairs $(\sigma_{i_k}, \sigma_{j_k})$ such that

$$(\mathrm{idx}_0(\sigma_{i_k}) - \mathrm{idx}_0(\sigma_{j_k}))(\mathrm{idx}_1(\sigma_{i_k}) - \mathrm{idx}_1(\sigma_{j_k})) < 0\,.$$

That is, σ_{i_k} and σ_{j_k} have different relative orders in idx_0 and idx_1. Let $\zeta_0 := \mathrm{idx}_0$, and for $k = 1, \ldots, m$, let $\zeta_k : \mathcal{K} \to \{1, \ldots, N\}$ be the simplex indexing obtained by transposing $(\sigma_{i_k}, \sigma_{j_k})$ in the simplex indexing ζ_{k-1}. If $\zeta_{k-1}(\sigma_{i_k})$ and $\zeta_{k-1}(\sigma_{j_k})$ are consecutive integers for all k, then $\zeta_m = \mathrm{idx}_1$. Furthermore, the sequence $\{(\sigma_{i_k}, \sigma_{j_k})\}_{k=1}^m$ can be ordered so that this conditions holds.

Proof. First, we prove that there is at least one pair $(\sigma_{i_k}, \sigma_{j_k})$ such that $\mathrm{idx}_0(\sigma_{i_k})$ and $\mathrm{idx}_0(\sigma_{j_k})$ are consecutive integers. Let $k_* = \arg\min_k \big|\mathrm{idx}_0(\sigma_{i_k}) - \mathrm{idx}_0(\sigma_{j_k})\big|$. To obtain a contradiction, suppose that $s_1 := \mathrm{idx}_0(\sigma_{i_{k_*}})$ and $s_2 := \mathrm{idx}_0(\sigma_{j_{k_*}})$ are not consecutive integers. Without loss of generality, $s_1 < s_2$. For $r = 1, \ldots, s_2 - s_1 - 1$, let $\tau_{s_1+r} := \mathrm{idx}_0^{-1}(s_1 + r)$ (i.e., $\tau_{s_1+1}, \ldots, \tau_{s_2-1}$ are the simplices between $\sigma_{i_{k_*}}$ and $\sigma_{j_{k_*}}$). For all r, we must have that either $(\sigma_{i_{k_*}}, \tau_{s_1+r}) \in \{(\sigma_{i_k}, \sigma_{j_k})\}_{k=1}^m$ or $(\sigma_{j_{k_*}}, \tau_{s_1+r}) \in \{(\sigma_{i_k}, \sigma_{j_k})\}_{k=1}^m$ (i.e., either the relative order of $\sigma_{i_{k_*}}$ and τ_{s_1+r} changes or the relative order of $\sigma_{j_{k_*}}$ and τ_{s_1+r} changes). By definition of k_*, we must have that none of the pairs in the set $\{(\tau_{s_1+r_1}, \tau_{s_1+r_2})\}_{r_1,r_2}$ are in $\{(\sigma_{i_k}, \sigma_{j_k})\}_{k=1}^m$ (i.e.,

the relative order of the simplices $\tau_{s_1+1}, \ldots, \tau_{s_2-1}$ does not change). Therefore we must either have $(\sigma_{i_{k_*}}, \tau_{s_1+1}) \in \{(\sigma_{i_k}, \sigma_{j_k})\}_{k=1}^m$ or $(\sigma_{j_{k_*}}, \tau_{s_1+r}) \in \{(\sigma_{i_k}, \sigma_{j_k})\}_{k=1}^m$ for all r. In either case, one of these is a transposition of simplices whose indices in idx_0 are consecutive integers, which is a contradiction.

Now we prove the lemma by induction on m. When $m = 1$, we showed above that $\mathrm{idx}_0(\sigma_{i_1})$ and $\mathrm{idx}_0(\sigma_{j_1})$ are consecutive integers, so $\zeta_1 = \mathrm{idx}_1$. In the general case, we can assume $\mathrm{idx}_0(\sigma_{i_1})$ and $\mathrm{idx}_0(\sigma_{j_1})$ are consecutive integers without loss of generality. The set $\{(\sigma_{i_k}, \sigma_{j_k})\}_{k=2}^m$ is the set of pairs $(\sigma_{i_k}, \sigma_{j_k})$ such that

$$(\zeta_1(\sigma_{i_k}) - \zeta_1(\sigma_{j_k}))(\mathrm{idx}_1(\sigma_{i_k}) - \mathrm{idx}_1(\sigma_{j_k})) < 0\,.$$

That is, σ_{i_k} and σ_{j_k} have different relative orders in ζ_1 and idx_1 for all $k = 2, \ldots, m$. By induction, we can assume that the sequence $\{(\sigma_{i_k}, \sigma_{j_k})\}_{k=2}^m$ is ordered such that $\zeta_{k-1}(\sigma_{i_k})$ and $\zeta_{k-1}(\sigma_{j_k})$ are consecutive integers for $k = 2, \ldots, m$ and $\zeta_m = \mathrm{idx}_1$. □

4.A.2 Modifications to step 1: Computing the polygons

In Section 4.3.1.1, we maintained a dictionary $\mathcal{D}_1(T)$ for each triangle $T \in B$. The keys were pairs (σ, τ) such that $I(\sigma, \tau) \cap T$ was a line segment in T, and the value of (σ, τ) was the list $[v, w]$ of vertices in $\mathcal{A}(L)$ that were the endpoints of the line segment $I(\sigma, \tau) \cap T$.

Now we maintain two additional dictionaries $\mathcal{D}_2(T)$ and $\mathcal{D}_3(T)$ for each triangle $T \in B$. These dictionaries are initialized to be empty, and are updated as we traverse the edges of B. At any time in this process, the keys of $\mathcal{D}_2(T)$ are pairs (v, w) of vertices in $\mathcal{A}(L)$ such that

1. the line segment (v, w) has been detected in T,
2. the line segment (v, w) is not an edge of B.

The value of $\mathcal{D}_2(T)[(v, w)]$ is a list $[(\sigma_{i_1}, \sigma_{j_1}), \ldots, (\sigma_{i_m}, \sigma_{j_m})]$ of the simplex pairs that we

have found so far such that σ_{i_k} and σ_{j_k} swap along (v, w). The keys of $\mathcal{D}_3(T)$ are vertices $v \in \mathcal{A}(L)$ such that

1. vertex v has been detected along an edge e of triangle T,
2. there is a pair (σ, τ) of simplices such that (σ, τ) swaps at v and we have not yet found a vertex w such that $I(\sigma, \tau) \cap T = (v, w)$.

The algorithm of Section 4.3.1.1 is modified as follows. Suppose that we detect a vertex v along edge e in $\mathcal{A}(L)$ for the set $\{(\sigma_{i_1}, \sigma_{j_1}), \ldots, (\sigma_{i_m}, \sigma_{j_m})\}$ of simplex pairs. We do the following:

1. **Update $\mathcal{D}_1$:** For each triangle $T \in B$ that is adjacent to e, we append v to the list of vertices for $\mathcal{D}_1(T)[(\sigma_{i_k}, \sigma_{j_k})]$ for all k, as in Section 4.3.1.1.

2. **Update $\mathcal{A}(L)$:** If v is not an endpoint of e, we split the edge e in $\mathcal{A}(L)$ and add an internal vertex in e, as in Section 4.3.1.1. If v is an endpoint of e, we do not split the edge or create a new vertex because B already has a vertex at v.

3. **Update $\mathcal{D}_2$, $\mathcal{D}_3$, and edge labels:** For each triangle T adjacent to e and each $(\sigma_{i_k}, \sigma_{j_k})$:

   - If v is the only vertex in the list $\mathcal{D}_1(T)[(\sigma_{i_k}, \sigma_{j_k})]$, then we have not yet detected a line segment for $(\sigma_{i_k}, \sigma_{j_k})$ of the form (v, w) for some vertex w. We do the following: If v is not in $\mathcal{D}_3(T)$, add key v to $\mathcal{D}_3(T)$ with value $[(\sigma_{i_k}, \sigma_{j_k})]$. Otherwise, append $(\sigma_{i_k}, \sigma_{j_k})$ to $\mathcal{D}_3(T)[v]$.
   - Otherwise, there is another vertex $w \in \mathcal{D}_1(T)[(\sigma_{i_k}, \sigma_{j_k})]$. This implies that we have just detected a line segment (v, w) in T for $(\sigma_{i_k}, \sigma_{j_k})$. We remove $(\sigma_{i_k}, \sigma_{j_k})$ from $\mathcal{D}_3(T)[w]$.

- If v and w are not both vertices of B, then (v, w) is not an edge of B. We do the following: If (v, w) is not in $\mathcal{D}_2(T)$, then add key (v, w) to $\mathcal{D}_2(T)$ with value $[(\sigma_{i_k}, \sigma_{j_k})]$. Otherwise, append $(\sigma_{i_k}, \sigma_{j_k})$ to $\mathcal{D}_2(T)[(v, w)]$.
- Otherwise, v and w are both vertices of triangle T. This means we have detected a line segment (v, w) in T in which v and w are the endpoints of an edge e' in B. If e' is an edge on the boundary of B, then we do nothing. Otherwise, let T_2 be the other triangle adjacent to e'. By Lemma 4.A.3, the pair $(\sigma_{i_k}, \sigma_{j_k})$ swaps along (v, w) if and only if the line segment is not detected in T_2. If e' already stores a reference to $(\sigma_{i_k}, \sigma_{j_k})$, then we remove it because this implies that e' was detected in T_2 already. Otherwise, we add a reference to $(\sigma_{i_k}, \sigma_{j_k})$ on e'.

When the traversal of the 1-skeleton is done, we add lines to $\mathcal{A}(L)$. For every triangle $T \in B$ and every key $(v, w) \in \mathcal{D}_2(T)$, we add a line segment with endpoints v, w to the DCEL that represents $\mathcal{A}(L)$. For every edge in the DCEL that is a subset of the line segment (v, w), we label the edge with a reference to the list $\mathcal{D}_2(T)[(v, w)]$, which is the list $\{(\sigma_{i_1}, \sigma_{j_1}), \ldots, (\sigma_{i_m}, \sigma_{j_m})\}$ of simplex pairs that swap along the line segment.

4.A.3 Modifications to step 2: Computing the pairing function

We compute a path Γ as in Section 4.3.1.2 and traverse Γ. At each step, we walk from one polygon P_1 to the next polygon P_2 by crossing an edge e in $\mathcal{A}(L)$. The edge e stores a list of simplex pairs (σ, τ) such that σ and τ have different relative orders in the polygons P_1, P_2. We update the simplex indexing one transposition at a time. Let $\overline{\text{idx}} : \mathcal{K} \to \{1, \ldots, N\}$ denote the current indexing, which we initialize to the simplex indexing $\text{idx}_f(\cdot, P_1)$ in P_1. While the list that e stores is nonempty, we do the following:

1. Let (σ, τ) be the first element of the list.
2. If $\overline{\text{idx}}(\sigma)$ and $\overline{\text{idx}}(\tau)$ are consecutive integers, then we update $\overline{\text{idx}}$ by swapping the

order of σ and τ. As in Section 4.3.1.2, we also update the RU decomposition, and the (birth, death) simplex pairs. We remove (σ, τ) from the list.

3. Otherwise, we move (σ, τ) to the end of the list.

At the end of this algorithm, $\overline{\text{idx}}$ is the simplex indexing $\text{idx}_f(\cdot, P_2)$ in P_2 (by Lemma 4.A.4), the RU decomposition is an RU decomposition for P_2, and we have computed the (birth, death) simplex pairs for P_2.

Acknowledgments

I thank Michael Lesnick and Nina Otter for helpful discussions.

CHAPTER 5

Analysis of Spatial and Spatiotemporal Anomalies Using Persistent Homology: Case Studies with COVID-19 Data

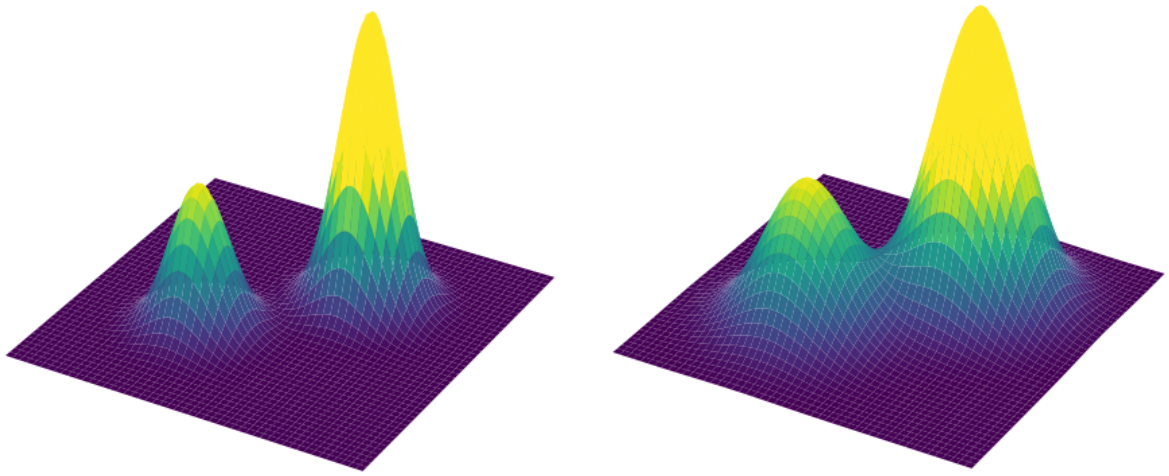
This chapter is primarily adapted from [HNP22], a paper that I led.[1] Appendix 5.B.3 is adapted from [FHP22].

5.1 Introduction

Many systems are spatial in nature. When working with spatial data sets, it is important to study the role of underlying spatial relationships [CG13]. To illustrate this importance, consider the spatiotemporal dynamics of Coronavirus disease 2019 (COVID-19) case rates. The spatial adjacencies between the neighborhoods of a city affect the dynamics of disease spread [PG16], and it is important to account for them. Researchers have studied a wide variety of spatial data sets, such as gross domestic product and life expectancy by country [Gap, BZ18] and voting in elections across different regions of a state [FP21]. Such data sets often also include temporal information (e.g., daily COVID-19 case rates), and it is also important to account for it.

We develop new methods for using TDA to analyze geospatial and geospatiotemporal

[1] I developed the methods, wrote the code, analyzed the results, and wrote the paper (with input and comments from Mason Porter and Deanna Needell).



(a) (b)

Figure 5.1.1: (a) The graph of a function $f : \mathbb{R}^2 \to \mathbb{R}$ that has two "well-separated" local maxima. (b) The graph of a function $g : \mathbb{R}^2 \to \mathbb{R}$ whose two local maxima have the same locations and values as f, but which are not well-separated from each other.

data sets in a way that directly incorporates spatial information. In this chapter, we treat geographical data as 2D data and construct a 2D filtered complex to represent it. Other spatial applications that have been examined using PH include sensor networks [SG07], percolation [SHC18], city-street networks [FP20], and the accessibility of polling places (see Chapter 6).

5.1.1 Our Contributions

We use TDA to analyze local extrema of real-valued geospatial data.[2] Our approach captures both local information (specifically, the geographical locations and the values of the local extrema) and global information about the relationships between the extrema. The global information includes the extent to which extrema are "spatially separated" (see Figure 5.1.1).

[2]See Section 5.3 for our definition of a "local maximum" and a "local minimum" of a real-valued function on a discrete set of geographical regions.

To the best of our knowledge, existing methods of analyzing local extrema yield only local information. One can check whether a geographical region is an extremum by comparing its value to those of its neighboring regions. However, this approach does not provide any global information about the extrema. For example, it cannot distinguish between the two cases in Figure 5.1.1.

When we examine time-dependent data, we use vineyards (see Section 2.5) to incorporate temporal information. Examining vineyards allows us to measure the persistence of extrema with time, observe how spatial separations between extrema change with time, and track how geographical locations of extrema change with time. We accomplish the last of these by using the vineyards to match the extrema at one time to the corresponding extrema at another time. (They may not be at the same geographical locations.) We identify the geographical locations of extrema by examining the sequence of (birth simplex, death simplex) pairs for each vine; the death simplex corresponds to the geographical location. Notably, we use information about the sequence of (birth simplex, death simplex) pairs for each vine, rather than using only the (birth, death) filtration values for each vine; to the best of our knowledge, our paper [HNP22], from which this chapter is adapted, is the first to do so. A naive approach, such as comparing each region to its neighboring regions at each time step, does not come with a natural way to match the extrema that one identifies at different times and does not provide information about changes in global structure. With our approach, we are able to track how the global spatial structure of data changes with time.

Another contribution of our research is a new method to construct an "efficient" simplicial complex whose underlying space[3] is homeomorphic to a geographical space (which is the set of regions, as we will explain shortly).[4] In our applications, we possess geographical data in the form of SHAPEFILEs. Each geographical region (e.g., a neighborhood or zip

[3]The *underlying space* of a simplicial complex is the union of its simplices. It is common in studies of TDA for authors to conflate the combinatorial and topological structures of a simplicial complex.

[4]The simplicial complex is "efficient" in the sense that it minimizes the number of simplices.

code) is represented in a SHAPEFILE by a polygon (or by multiple polygons, if the region is disconnected) with many vertices (about 100 to 1000 vertices, depending on the particular SHAPEFILE and the particular region). These polygons approximate the real-life boundaries of the geographical regions. A naive approach to building a simplicial complex is to simply triangulate each of the polygons. However, this approach has two issues. The first is that there are often small overlaps between the polygons or spurious gaps between the polygons because the polygon boundaries do not exactly match the real-life geographical boundaries. The vertices of a polygon often lie in the interior of another polygon. The second issue is that simply triangulating these polygons, which each have a very large number of vertices, would create orders-of-magnitude more simplices than are necessary to represent a geographical space. It is important to attempt to minimize the number of simplices in a simplicial complex because PH and vineyard computation times are very sensitive to the number of simplices.

Rather than naively triangulate the given polygons, we use the SHAPEFILE of a geographical space to infer adjacency information about the associated regions; we then use only this information to build a simplicial complex for that geographical space. In the resulting simplicial complex, each region is represented by a union of triangles. We use about 1 to 10 triangles per region, depending on the number of neighbors of the region. By contrast, the naive approach above requires about 100 to 1000 triangles per region. Two adjacent regions that have a connected intersection share exactly one edge in our simplicial complex, except in rare special cases that we will discuss in Section 5.2. In our simplicial complex for a geographical space, the union of any subset of geographical regions is homeomorphic to the underlying space of the simplicial subcomplex (see Section 5.2 for the definition of a simplicial subcomplex) that is induced by the union of the corresponding triangles. When the geographical regions satisfy the mild assumptions (A1)–(A4) that we define in Section 5.2, our construction uses the minimum number of simplices that is possible for a simplicial complex with the property above. (See Property (P) in Section 5.2.)

As case studies, we apply our approach to two data sets. The first data set is a geospatial

data set of per capita vaccination rates in New York City (NYC) by zip code [Cit]. The homology classes correspond to zip codes in which the vaccination rate is either lower or higher (depending on choices that one can make in our approach) than in the neighboring zip codes. The estimates of these rates are at a single point in time (23 February 2021). The second data set consists of 14-day mean per capita COVID-19 case rates in neighborhoods in the city of Los Angeles (LA) in the time period 25 April 2020–25 April 2021. Modeling the spatiotemporal spread of COVID-19 is a complex task [Ari21, VTD20]. In this geospatiotemporal data set, the homology classes of our approach correspond to COVID-19 anomalies, which are regions whose case rates are higher than in the neighboring regions.[5] It is important to examine such anomalies, as COVID-19 spreads with significant spatial heterogeneity and thus has heterogeneous effects on different areas.[6] Many factors (such as mobility, population density, socioeconomic differences, and racial demographics) play a role in how COVID-19 affects regions differently [HSA21, HGL21, Cen]. In our case study of COVID-19 case rates in LA, we construct a vineyard that (1) conveys which anomalies are most persistent in time and (2) reveals how the anomalies move geographically with time.

5.1.2 Related Work

Our method addresses several limitations of previous efforts to combine TDA with geospatial analysis. In [SHP16], Stolz, Harrington, and Porter studied the percentage of United Kingdom voters by district that voted to leave the European Union in the "Brexit" referendum. The holes that they identified using PH correspond to districts that voted differently than

[5] We examine *local* maxima in the case-rate data. This contrasts to COVID-19 "hotspots," which the CDC has defined using an absolute threshold for the number of cases and criteria that are related to the temporal increase in the number of cases [OKC20].

[6] Other scholars have studied contagions using TDA in ways that do not yield topological features with geographical meaning. For example, recent work used TDA to study the spatiotemporal spread of COVID-19 [SZW21] and Zika [SLG20]. These papers examined topological features in atmospheric data, which were then used to forecast case rates. TDA was also used in [TKH15] to study the Watts threshold model of a social contagion on noisy geometric networks.

the neighboring districts. However, their approach does not distinguish between homology classes that were merely noise and homology classes that correspond to small geographical districts. In [FP21], Feng and Porter developed an approach to study PH by constructing filtered complexes using the level-set method [OF03] of front propagation from scientific computation.[7] Using their level-set complexes, they examined the percentage of voters in each precinct of California counties that voted for a given candidate (e.g., Hillary Clinton) in the 2016 United States presidential election. The homology classes represent precincts that voted more heavily for Clinton than the neighboring precincts. The level-set complexes in [FP21] have two key limitations. The first is that they cannot handle time-dependent data, as they are built to study either data at a single point in time or data that has been aggregated over some time window to yield time-independent data. The second limitation is that these simplicial complexes reduce real-valued data (e.g., the percentage of voters who voted for Clinton) to binary data (e.g., whether or not the majority voted for Clinton). Consequently, in this example, the level-set-based PH does not capture the extent to which a blue "political island" voted more heavily for Clinton. By contrast, our approach is designed specifically to capture such information. As a trade-off, we no longer capture the geographical sizes of the political islands. For further discussion, see Feng, Hickok, and Porter [FHP22], who applied the level-set filtration to study the cumulative case count in Los Angeles on one specific day.

Our new approach to compute PH is also able to resolve some other technical issues in [FP21]. In particular, some of the homology classes in the level-set approach of [FP21] are geographical artifacts that are indistinguishable from true features of a data set. By contrast, the finite 1D homology classes in our approach are either in one-to-one correspondence with the local maxima of a real-valued geospatial function or in one-to-one correspondence with its local minima, depending on the choices that one makes. Additionally, unlike the level-set

[7]The name "level-set method" may cause confusion. Importantly, the level-set simplicial complex of [FP21] is not the simplicial subcomplex that has simplices with some prescribed filtration value (i.e., a level set of the filtration values of a simplicial complex).

approach in [FP21], we are able to detect extrema that are adjacent to the boundary of a geographical space.

Other methods to construct simplicial complexes from geospatial data, such as rasterization of a SHAPEFILE or treating the regions as a point cloud, require a trade-off between the number of simplices and the accuracy of the representation of the geographical regions. For example, the level-set-based PH method of [FP21] uses orders-of-magnitude more simplices to achieve sufficient resolution of the smallest geographical regions (e.g., densely populated urban centers that are important to analyze). See Section 5.6 for further discussion.

We use vineyards in the present chapter, but there are also other ways to study the topology of time-varying data. For example, zigzag PH [CS10] was used in [CJ17] to analyze time-dependent point clouds (such as swarms) and in [TMK20] to study time-delay embeddings of dynamical systems. Crocker plots and crocker stacks (i.e., stacks of crocker plots for different values of a smoothing parameter) illustrate how the Betti numbers of a time-dependent point cloud change with time and with a scale parameter r [XAT22]. Additionally, Kim and Mémoli [M21] used multiparameter PH [CZ07] to study time-dependent point clouds. In Appendices 5.B.4 and 5.B.5, we show how one can use multiparameter PH [CZ07, BL22] and multiparameter zigzag PH [CS10] to study our COVID-19 spatiotemporal data sets.

5.1.3 Organization

The remainder of this chapter proceeds as follows. In Section 5.2, we formulate how we construct simplicial complexes. In Section 5.3, we define several filtration functions and discuss how to interpret the resulting PDs and vineyards. In Section 5.4, we apply our method to the LA and NYC data sets. In Section 6.4, we discuss our methodological choices. In Section 5.6, we summarize our work and discuss some of its implications. In the appendix, we discuss technical details of the simplicial complex construction, discuss alternative topological approaches for studying PH in geospatiotemporal data,

provide further information about the LA results, compare our approach to an "all-but-one" statistical test, and show some demographic data. Our code is available at `https://bitbucket.org/ahickok/vineyard/src/main/`.

5.2 Constructing a Simplicial Complex

We now show how we construct a simplicial complex $\mathcal{K}$ from geographical data (e.g., a SHAPEFILE that specifies approximate geographical boundaries of a set of geographical regions). We partition a given geographical space into *regions*. In Section 5.4.1, the regions are zip codes in NYC; in Section 5.4.2, the regions are neighborhoods in the city of LA. Let S be the set of regions. We refer to the complement of $\bigcup_{R \in S} R$ as the *exterior region*. We construct a 2D simplicial complex $\mathcal{K}$ with the following property:

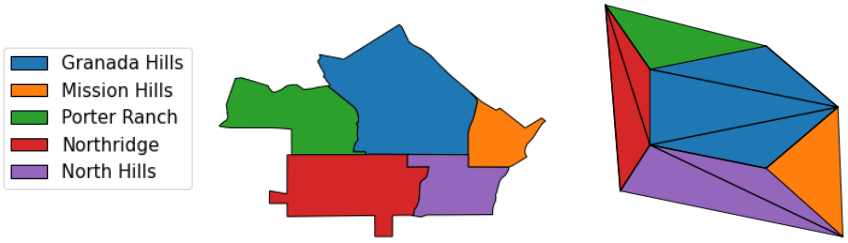
(P) There is an assignment of 2D simplices to regions such that the union of any subset of regions is homeomorphic to the underlying space of the *simplicial subcomplex*[8] that is induced by the union of the corresponding 2D simplices.

In Figure 5.2.1, we show an example of our construction, which we discuss in this section and present in more detail in Appendix 5.A. Under the mild assumptions (A1)–(A4) that we state shortly, our simplicial complex has the minimum number of simplices that is possible for a simplicial complex that satisfies property (P). Constructing an efficient simplicial complex is important because the time that it takes for TDA computations depends sensitively on the number of simplices in a simplicial complex.

In our case studies, the geographical data take the form of SHAPEFILEs. In a SHAPEFILE, each region is represented by a *polygon with holes*[9] (or by multiple polygons with holes, if

[8]The simplicial subcomplex that is induced by a set $E \subseteq \mathcal{K}$ is the smallest simplicial complex $\mathcal{K}'$ that contains the set E of simplices. That is, if $\mathcal{K}''$ is a simplicial complex that contains E, then $\mathcal{K}' \subseteq \mathcal{K}''$. When $\mathcal{K}$ is 1D, a simplicial subcomplex is equivalent to an induced subgraph.

[9]A polygon with holes is of the form $P = Q - \bigcup_{i=1}^{h} \text{int}(H_i)$, where Q is a polygon that encloses polygons



(a) (b)

Figure 5.2.1: (a) A set S of geographical regions, as given by a SHAPEFILE [GEO]. (b) The resulting simplicial complex $\mathcal{K}$.

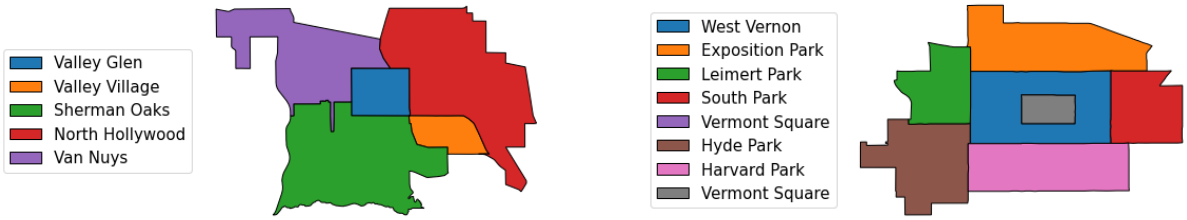
the region is disconnected) that closely approximates the actual geographical region. (A SHAPEFILE stores the coordinates of the boundaries of the polygons.) For an example of SHAPEFILE data, see Figure 5.2.1a. As we discussed in Section 5.1.1, the polygon boundaries are not always aligned perfectly, so their interiors sometimes overlap and gaps can occur between them. Therefore, to construct a simplicial complex $\mathcal{K}$, we must do more than merely triangulate these polygons. Additionally, the polygons in our SHAPEFILEs have roughly between 100 and 1000 vertices, which is many more vertices per region than in the simplicial complex $\mathcal{K}$ that we will construct shortly.

We make the following assumptions about geographical regions:

(A1) There are a finite number of regions, and each region has a finite number of connected components.

(A2) Each component of a region is homeomorphic to $D_0 - \bigcup_{i=1}^{h} \text{int}(D_i)$, where D_0 is a closed disk that encloses some number (which can be 0) of other closed disks $D_1, \ldots, D_h$ (i.e., the holes of the region). For all $i \neq j$, the intersection $D_i \cap D_j$ has at most one point. See, for example, the West Vernon region in Figure 5.2.2b; it is homeomorphic to $D_0 - D_1$ (an annulus) for two disks D_0 and D_1 that do not intersect. (In our case

$H_1, \ldots, H_h$ (the holes) [OR87] and $\text{int}(H_i)$ denotes the interior of H_i. It is possible to have $h = 0$ holes.



(a) Valley Glen (b) West Vernon

Figure 5.2.2: Various neighborhoods of Los Angeles, as given by a SHAPEFILE [Geo]. (a) The four neighborhoods Valley Glen, Valley Village, Sherman Oaks, and North Hollywood intersect in a point. (b) The neighborhood West Vernon has a hole because of its neighbor Vermont Square.

studies, it is rare for any of the disks to intersect.)

(A3) The intersection between any two regions has a finite number of components, and the interiors of the regions do not intersect.

(A4) The intersection between three or more regions is either a point or $\emptyset$.

Assumptions (A1)–(A4) are very reasonable for human-made geographical boundaries. We do not even require the regions to be simply connected or the region intersections to be connected. In Figure 5.2.1a, we illustrate the most typical situation that we encounter. In this example, LA neighborhood Granada Hills is homeomorphic to a disk and its boundary intersects the boundaries of five neighboring regions (counting the exterior region). In Figures 5.2.2 and 5.2.3a, we illustrate a few other configurations that can arise in geospatial applications.

We now outline our procedure for building a simplicial complex. For each region R, we construct a "reduced" polygon with holes P^R that has orders-of-magnitude fewer vertices than the polygons with holes in the associated SHAPEFILE. The number of holes in P^R is equal to the number of holes of the geographical region R. We glue the boundaries of

$\{P^R \mid R \in S\}$ together in a way that respects the geographical region boundaries. We then triangulate each of the polygons to obtain a 2D simplicial complex $\mathcal{K}$. We assign a 2D simplex $\sigma \in \mathcal{K}$ to the region R whose polygon P^R originally contained σ. In Figure 5.2.1, we show an example of the resulting $\mathcal{K}$. Our code for our simplicial-complex algorithm is available at `https://bitbucket.org/ahickok/vineyard/src/main/`.[10] In the remainder of this section, we discuss the details of this process.

5.2.1 Constructing a Reduced Polygon with Holes for each Region

Without loss of generality, we assume that each region is connected; if not, we treat each component of a region as if it were its own region. For each region R, we construct a reduced polygon with holes P^R using only adjacency information that we infer from a SHAPEFILE. Let $D_0, D_1, \ldots, D_h$ be the disks in the statement of assumption (A2), and let $B_i = \partial D_i$. Under the geographical assumptions (A1)–(A4), the intersections of a region R with its neighbors are such that for each boundary B_i, one can order the neighbors in clockwise (or counterclockwise) fashion, possibly with repetition.[11] Let S_i denote this sequence of neighbors around B_i. We list intersections with the exterior region in the same manner as for any other neighboring region. We also record whether each intersection is 1D or 0D. For example, in Figure 5.2.2a, the clockwise sequence of neighbors around the boundary of Valley Glen is {Van Nuys, North Hollywood, Valley Village, Sherman Oaks}. The intersection with Valley Village is 0D and the other intersections are 1D. For regions such as West Vernon in

[10]This code has one limitation that the algorithm in the present chapter does not. It requires that no interior region (i.e., a region that is contained in the outer boundary of another region) intersects any other interior region. This does not occur in our data, and we believe that it does not occur in most geographical spaces.

[11]Theoretically, several 0D intersections can be adjacent to each other, although this scenario does not occur in our data sets. That is, in principle, there can exist a sequence $\{N_i, \ldots, N_{i+k}\}$ of neighbors such that $N_j \cap R$ is the same point p for all j. The order of this sequence is not determined uniquely by the intersections of the neighbors with R. Instead, we order them in the order in which they appear clockwise (or counterclockwise) around the point p. This sequence must be finite because there are a finite number of regions and (A2) implies that $N_{j_1} \neq N_{j_2}$ if $j_1 \neq j_2$.

Figure 5.2.2b, we obtain a sequence S_i for each boundary B_i. Each sequence is unique up to the choice of starting neighbor.

Given a sequence of neighbors for each boundary B_i (which, if necessary, we adjust as in Appendix 5.A.1), we construct a polygon with holes P^R as follows. Let $(P')^R$ be a polygon that has one edge for each $N \in S_0$ for which the corresponding component of $N \cap B_0$ is 1D. Let $\{H_i^R\}_{i=1}^h$ be a set of polygons that are contained in $(P')^R$ and satisfy the following properties:

1. H_i^R has one edge for each $N \in S_i$ for which the corresponding component of $N \cap B_i$ is 1D,
2. $H_i^R \cap H_j^R \neq \emptyset$ if and only if $D_i \cap D_j \neq \emptyset$,
3. $P^R \cap H_i^R \neq \emptyset$ if and only if $D_0 \cap D_i \neq \emptyset$, and
4. if the intersection of two polygons in $\{P^R, H_1^R, \ldots, H_h^R\}$ is nonempty, then the intersection is a vertex.

The locations of the vertices do not matter. We define P^R to be $(P')^R - \bigcup_{i=1}^h \text{int}(H_i^R)$, which is homeomorphic to R by assumption (A2). Finally, we annotate each edge of P^R with the neighbor that corresponds to it. We also annotate each vertex with the sequence of its adjacent regions, which we list in clockwise order starting with R.

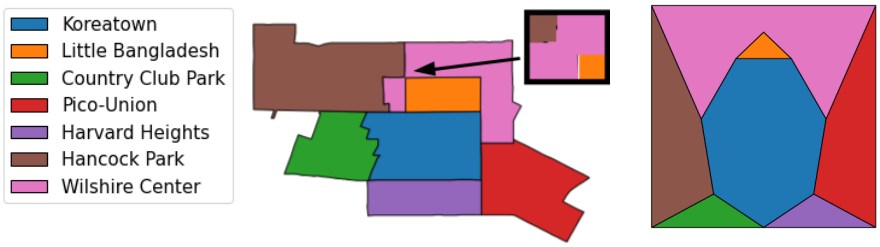
5.2.2 Gluing Together the Polygons with Holes

We glue the polygons with holes $\{P^R \mid R \in S\}$ along their edges according to their edge and vertex annotations. More precisely, if P^{R_1} has n nonadjacent edges with the annotation R_2 (which is the typical situation when $R_1 \cap R_2$ has n components that are 1D), then P^{R_2} has exactly n nonadjacent edges with the annotation R_1. For example, in Figure 5.2.3, $R_1 =$ Koreatown and the annotated polygon with holes P^{R_1} has two edges with the annotation $R_2 =$ Wilshire Center. Let (u, v), with u and v in clockwise order, be the vertices of an

edge in P^{R_1} with annotation R_2. Because the n edges are nonadjacent, u and v must each have at least three neighbors (including R_1 and R_2). For example, in Figure 5.2.3, again consider the two edges with the annotation Wilshire Center. The two vertices u_1 and v_1 of one edge have the adjacency sequences {Koreatown, Hancock Park, Wilshire Center} and {Koreatown, Wilshire Center, Little Bangladesh}, respectively. The two vertices u_2 and v_2 of the other edge have the adjacency sequences {Koreatown, Little Bangladesh, Wilshire Center} and {Koreatown, Wilshire Center, Pico-Union}, respectively. For a given (u, v), we seek an edge (x, y) (with x and y in clockwise order) in P^{R_2} with the annotation R_1 such that (1) u and y are annotated with the same sequences and (2) v and x are annotated with the same sequences. We know that there must be at least one such edge because (u, v) represents a component of $R_1 \cap R_2$ and there is some edge in P^{R_2} that represents the same component (so its vertices have the same sequences of adjacent regions as u and v). In Lemma 5.A.2, we prove that there is a unique such edge. If there are $n > 1$ consecutive edges $e_0, \ldots, e_{n-1}$ on the boundary of $\mathcal{K}_{R_1}$ with annotation R_2, then there are n consecutive edges $e'_0, \ldots, e'_{n-1}$ on the boundary of $\mathcal{K}_{R_2}$ with annotation R_1. This situation arises precisely because of the adjustments that we discuss in Appendix 5.A.1. We glue e_i to e'_{n-i} for all i. If $R_1 \cap R_2$ is homeomorphic to S^1, then the choice of e'_0 as the first edge in P^{R_2} is not unique, but all choices result in topologically equivalent spaces. In Figure 5.2.3b, we show the result of the gluing process for Koreatown and its neighbors.

5.2.3 Triangulating the Polygons with Holes

We triangulate each polygon with holes P^R using the inductive algorithm in [OR87]. We show examples of triangulated polygons with holes in Figure 5.2.4. The result of this triangulation process is a 2D simplicial complex $\mathcal{K}$ with property (P). (We assign a 2D simplex in the polygon with holes P^R to the geographical region R.) The simplicial complex $\mathcal{K}$ is a minimal simplicial complex with property (P) because (1) each polygon with holes P^R has the minimum number of vertices and holes and (2) the number of triangles in the triangulation



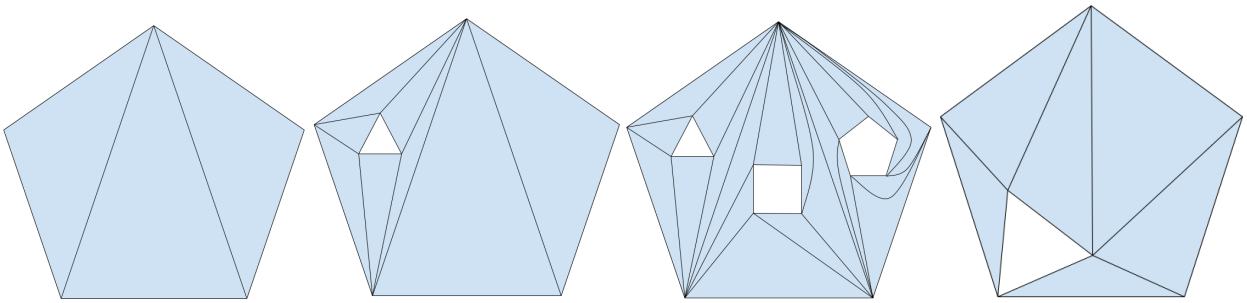
(a) (b)

Figure 5.2.3: (a) A geographical set S that consists of the neighborhood Koreatown and its neighbors, as given by a SHAPEFILE [Geo]. Observe that the neighborhood Little Bangladesh has only two neighbors and that the intersection between Koreatown and Wilshire Center has two components. (b) The result of gluing Koreatown's polygon to the polygons of its neighbors.

of P^R is determined by its number of vertices and its number of holes by Euler's theorem (see [OR87]). For an example of a triangulated polygon with holes, see Figure 5.2.1b.

5.3 Our Filtration Functions

We define various filtrations that one can use with the simplicial complex $\mathcal{K}$ that we constructed in Section 5.2, and we discuss how to interpret the resulting PDs and vineyards. Let S be the set of geographical regions R that the simplicial complex $\mathcal{K}$ represents, and let $F : S \to \mathbb{R}$ be a real-valued function on S. For example, in Section 5.4.1, $F(R)$ is the per capita full-vaccination rate (i.e., having received all required doses of some vaccine) for COVID-19 in NYC zip code R. In Sections 5.3.1 and 5.3.2, we define two filtration functions that are induced by F. Given a time-dependent and real-valued function $F(t, R)$, we define time-dependent filtration functions in Section 5.3.3. For example, in Section 5.4.2, $F(t, R)$ is the 14-day mean per capita COVID-19 case rate in neighborhood R on day t. From a time-dependent filtration function, we compute a vineyard.



(a) (b) (c) (d)

Figure 5.2.4: Triangulation of a polygon with holes P^R for a region R when (a) R has no holes, (b) R has a single hole, (c) R has multiple holes, and (d) R has a hole that touches the exterior boundary of R.

5.3.1 The Sublevel-Set Filtration

In this subsection, we define a sublevel-set filtration. In our applications, we use the 1D PH of the sublevel-set filtration to analyze local maxima in our data sets. We illustrate the idea of a sublevel-set filtration in Figure 5.3.1.

Definition 5.3.1 (Sublevel-set filtration)**.** Let $\mathcal{K}$ be the simplicial complex that we obtain from our construction in Section 5.2 for a set S of regions, and let g be the assignment of 2D simplices to the regions. Let $F : S \to \mathbb{R}$. We define the sublevel-set filtration function f by considering the sublevel sets of F. On the 2D simplices σ, we define the filtration function by

$$f(\sigma) = F(g(\sigma)) \,.$$

We extend the filtration function to the remaining (lower-dimensional) simplices as follows. If σ is a vertex or edge on the boundary of $\mathcal{K}$, we set

$$f(\sigma) = \min_R F(R).$$

Otherwise, we set

$$f(\sigma) = \min\{f(\tilde{\sigma}) \mid \tilde{\sigma} \text{ is a 2D simplex for which } \sigma \text{ is a vertex or edge of } \tilde{\sigma}\}. \tag{5.1}$$

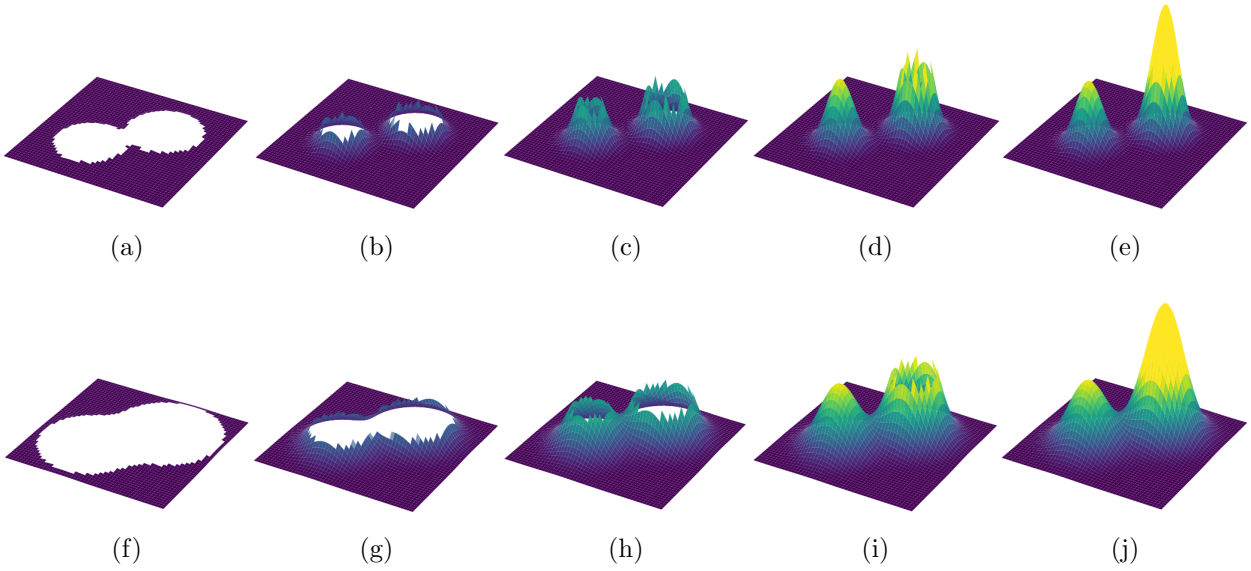


Figure 5.3.1: In panels (a)–(e), we show the α-sublevel sets for increasing α of a function $f : \mathbb{R}^2 \to \mathbb{R}$ that has two well-separated local maxima. In (a), for the smallest value of α, there is one hole that corresponds to the global maximum. In (b), a second hole appears; it corresponds to the other local maximum. In (d), the second hole is filled in. In (e), the first hole is filled in. In panels (f)–(j), we show the α-sublevel sets for increasing α of a function $g : \mathbb{R}^2 \to \mathbb{R}$ whose two local maxima have the same locations and values as f, but are not well-separated from each other. The second hole does not appear until the sublevel set in panel (h). In all panels, the jagged edges are artifacts of the way that the Python package MATPLOTLIB plots surfaces.

At filtration parameter-value α, the simplicial complex $\mathcal{K}_\alpha$ is the simplicial subcomplex of $\mathcal{K}$ that is induced by the union of the set of 2D simplices σ such that $F(g(\sigma)) \leq \alpha$ and the set of vertices and edges that are on the boundary of $\mathcal{K}$. Henceforth, we say that the vertices and edges on the boundary of $\mathcal{K}$ are "exterior-adjacent." By construction, the underlying space of $\mathcal{K}_\alpha$ is homeomorphic to the union of all regions R such that $F(R) \leq \alpha$ and the exterior boundary. We set $f(\sigma) = \min_R F(R)$ for exterior-adjacent vertices and edges σ for technical reasons that we will explain in a few paragraphs. In appendix 5.B.2, we explore an alternative definition in which we set the filtration values of exterior-adjacent vertices and edges σ to $\min_R\{F(R) \mid R \subseteq C\}$, where C is the connected component that contains σ.

The 1D PH of the sublevel-set filtration encodes information about the structure of the local maxima of F. A region R of a geographical space is a local maximum if the value of $F(R)$ is larger than the value of $F(N)$ for all neighboring regions N of R for which $N \cap R$ is 1D. More generally, we consider a set $E \subseteq S$ of regions (where $|E| = 1$ is possible) to be a *local maximum* if

1. the interior of $\bigcup_{R \in E} E$ is connected,
2. the value of F is constant on E (we denote this value by $F(E)$), and
3. the value $F(E)$ is larger than the value $F(N)$ for all regions $N \notin E$ such that $N \cap R$ is 1D for some $R \in E$.

If E is a local maximum, there is a 1D homology class whose death simplex is one of the simplices in the preimage $g^{-1}(E)$, where g is the map from $2D$ simplices in $\mathcal{K}$ to geographical regions in S. The class dies at filtration parameter-value $\alpha = F(E)$. For example, if $F(R)$ is the COVID-19 case rate in region R, then the 1D homology classes correspond to COVID-19 anomalies and the death simplex of a 1D homology class indicates the epicenter of that anomaly. The larger the value $F(E)$ in comparison to nearby regions (including regions that are not necessarily immediate neighbors), the more persistent the homology class is. If the

union of all regions (excluding the exterior region) is not simply connected, then there is at least one 1D homology class with an infinite death time. See Figure 5.4.1b for an example. The infinite 1D homology classes correspond to the holes in the geographical space, rather than to local maxima. The local maxima of F are in one-to-one correspondence with the set of 1D homology classes with finite death times.[12] There is a canonical mapping from finite 1D homology classes to regions. A class that is represented by the simplex pair (σ_b, σ_d) is mapped to the region $g(\sigma_d)$ that includes σ_d. The region $g(\sigma_d)$ is the location of the local maximum of F that corresponds to the homology class,[13] and the death simplex's filtration value $f(\sigma_d)$ is the value of the local maximum. The death simplices of the finite 1D homology classes and their filtration values give the local-maximum locations R and their function values $F(R)$.

With the 1D PH, we can do more than simply identify local maxima and their locations; the 1D PH also reveals information about relationships between the local maxima. If the local maxima are well-separated from one another, then the corresponding homology classes all have early birth times. For example, the NYC data set has several connected components. One can think of the global maximum of each connected component as being "totally separated" from each other because they are on different connected components. The corresponding 1D homology classes are all born at the earliest possible filtration time, which is $\min_R F(R)$ (see Figure 5.4.2a). We show an example of well-separated local maxima in Figure 5.1.1a. By contrast, the two local maxima in Figure 5.1.1b are not well-separated, so the homology class that corresponds to the lower peak in Figure 5.1.1b is born at a larger filtration value than the homology class in Figure 5.1.1a. See Figure 5.3.1 for a visualization

[12]Recall that in our definition of a local maximum, we only compare the value of a region R (or the constant value of a set E of regions) to the values of neighbors N that have a 1D intersection with R (or with a region in E). It is possible for two local maxima, R_1 and R_2, to have a 0D intersection. In that case, let N be the set of regions that are adjacent to $R_1 \cup R_2$. Then N is homotopy-equivalent to a figure-8, which has two 1D homology generators. One of them corresponds to R_1, and the other one corresponds to R_2.

[13]Let $E \subseteq S$ be the local maximum that corresponds to the 1D homology class. If $E = \{R\}$, then $g(\sigma_d) = R$. However, if E contains multiple regions, then $g(\sigma_d)$ is only one of the regions in E.

of the sublevel sets. The birth times of the 1D homology classes reflect structural information about the local maxima.

We set the filtration value of exterior-adjacent vertices and edges to the global minimum $\min_R F(R)$ so that 1D PH can detect local maxima on the boundary of a geographical space. (We consider an alternative approach in appendix 5.B.2.) This is important for the LA data set of COVID-19 case rates. In Figure 5.4.7, we observe that many of the most-persistent COVID-19 anomalies are on the boundary of the geographical space; it is crucial that we are able to detect them. If we had not defined the exterior-adjacent filtration values in this way, then the filtration value of exterior-adjacent vertices and edges σ would be $F(R)$, where R is the unique region that is adjacent to σ. If R is a local maximum, its corresponding 1D homology class is born and dies at filtration parameter-value $\alpha = F(R)$. In the PD, it then appears as a point on the diagonal. Therefore, for 1D PH to detect local maxima on the boundary of a geographical space, we must adjust the filtration values of exterior-adjacent vertices and edges.

The 0D homology classes correspond to local minima of F. However, unlike for the 1D homology classes, there is not a natural mapping from 0D homology classes to the locations of the minima. In appendix 5.B.1, we discuss the interpretation and computation of 0D homology classes.

5.3.2 The Superlevel-Set Filtration

An alternative to using the sublevel-set filtration from Section 5.3.1 is to use superlevel sets of F to construct a superlevel-set filtration. In our case study on COVID-19 vaccination rates in NYC, we use a superlevel-set filtration to analyze local minima of the vaccination rate. We define a *local minimum* analogously to the way that we defined a local maximum in Section 5.3.1. We illustrate the idea of the superlevel-set filtration in Figure 5.3.2.

Definition 5.3.2 (Superlevel-set filtration)**.** Let $F : S \to \mathbb{R}$ for a set S of regions. The

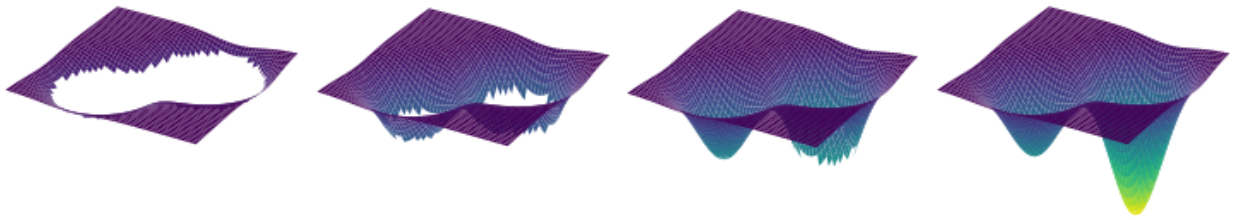


Figure 5.3.2: The α-superlevel sets, with α decreasing from left to right, for the graph of a function $f : \mathbb{R}^2 \to \mathbb{R}$ with two local minima.

superlevel-set filtration function f is the sublevel-set filtration function that is induced by $-F$.

At filtration parameter-value α, the simplicial complex $\mathcal{K}_{-\alpha}$ is the simplicial subcomplex of $\mathcal{K}$ that is induced by the union of the set of exterior-adjacent simplices and the set of 2D simplices σ for which $F(g(\sigma)) \geq \alpha$. By construction, the underlying space of $\mathcal{K}_{-\alpha}$ is homeomorphic to the union of regions R for which $F(R) \geq \alpha$ along with the exterior boundary. Local maxima of F now correspond to 0D homology classes, and local minima of F now correspond to 1D homology classes; this is the opposite situation from the sublevel-set filtration. Our discussion of local maxima for the sublevel-set filtration in Section 5.3.1 applies to local minima for the superlevel-set filtration, and our discussion of local minima for the sublevel-set filtration in Section 5.3.1 applies to local maxima for the superlevel-set filtration. The only difference is that the filtration values in the superlevel-set filtration are the additive inverses of the function values of F. This implies, for example, that the death filtration value of a 1D homology class that corresponds to a local minimum at region R is $\alpha = -F(R)$, rather than $\alpha = F(R)$.

5.3.3 A Time-Dependent Filtration

Suppose that we have a time-dependent, real-valued function $F(t, R)$ whose domain is $\{t_0, t_1, \ldots, t_n\} \times S$, where $t_0 \in \mathbb{R}$ is the initial time and $t_n \in \mathbb{R}$ is the final time. For

example, in Section 5.4.2, the value $F(t, R)$ is the 14-day mean per capita COVID-19 case rate in neighborhood R on day t. We seek to analyze the structure of local extrema as they change with time.

Definition 5.3.3 (Time-dependent sublevel-set filtration)**.** Let $F : \{t_0, t_1, \ldots, t_n\} \times S \to \mathbb{R}$ be a time-dependent function on a set S of regions, and let $\mathcal{K}$ be the simplicial complex for S from the construction in Section 5.2. At each time $t_i \in \{t_0, t_1, \ldots, t_n\}$, we define the time-dependent filtration function $f(t_i, \cdot)$ to be the sublevel-set filtration that is induced by $F(t_i, \cdot)$. To extend this filtration function to the entire interval $[t_0, t_n]$, we linearly interpolate $f(\cdot, \sigma)$ on each subinterval $[t_i, t_{i+1}]$ for all simplices $\sigma \in \mathcal{K}$.

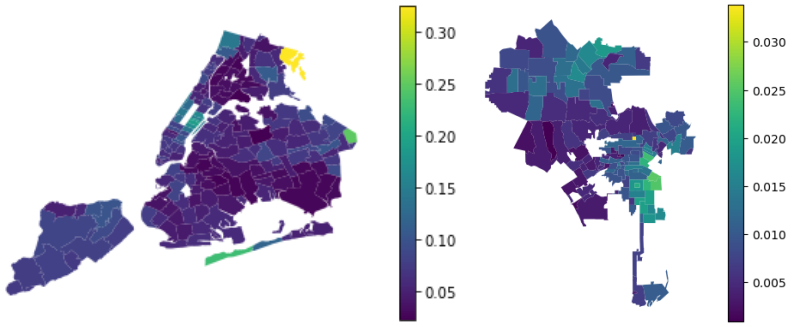
In the present chapter, we only use the time-dependent sublevel-set filtration, but one can analogously define a time-dependent superlevel-set filtration. We have implemented both of these filtrations in our code.

We use a time-dependent sublevel-set filtration to construct a vineyard. This allows us to track how the extrema move in both space and time. As in Section 5.3.1, each finite vine corresponds to a local maximum whose location at time t is given by the region $g(\sigma_d(t))$ that contains the vine's time-dependent death simplex $\sigma_d(t)$.[14]. The length of a vine corresponds to its persistence in time.

5.4 Case Studies

We now apply our methods to two data sets, which we illustrate in Figure 5.4.1.

[14]It is known that vineyards are not stable; see Appendix 3.A in Chapter 3 for an example of vineyard instability. A small perturbation in filtration values can cause crossing of vines that previously did not cross (i.e., it is an "avoided crossing") This, in turn, causes simplex pairings to change. Therefore, the geographical region $g(\sigma_d(t))$ that corresponds to a particular vine at time t is sensitive to small perturbations in filtration values.



(a) NYC zip codes (b) LA neighborhoods

Figure 5.4.1: We show (a) the per capita COVID-19 full vaccination rate in NYC by (modified) zip code on 23 February 2021 and (b) the 14-day mean per capita COVID-19 case rate in the city of LA by neighborhood on 30 June 2020. In both (a) and (b), the white regions are geographical regions that do not belong to the depicted city.

5.4.1 COVID-19 Vaccination Rates in New York City

We examine vaccination rates in (modified) zip codes of NYC.[15] We demonstrate the effects of the two filtrations that we defined in Section 5.3. The geographical boundaries of the zip codes are given by a SHAPEFILE [NYCc]. From the SHAPEFILE, we construct a simplicial complex $\mathcal{K}$ in the manner that we described in Section 5.2. The vaccination data set, which we obtained from the NYC Department of Health & Mental Hygiene website [Cit], consists of the number of fully vaccinated people in each zip code on 23 February 2021.[16] For each zip code, we divide this number by its population estimate in [Cit] to obtain a per capita vaccination rate. For zip code R, we define $F(R)$ to be the per capita vaccination rate in R

[15]The NYC Department of Health and Mental Hygiene uses modified zip code tabulation areas (MODZCTA) for COVID-19 data [NYCc]. In these modified zip codes, some zip codes with small populations are combined [NYCb]. We henceforth refer to modified zip codes as simply "zip codes."

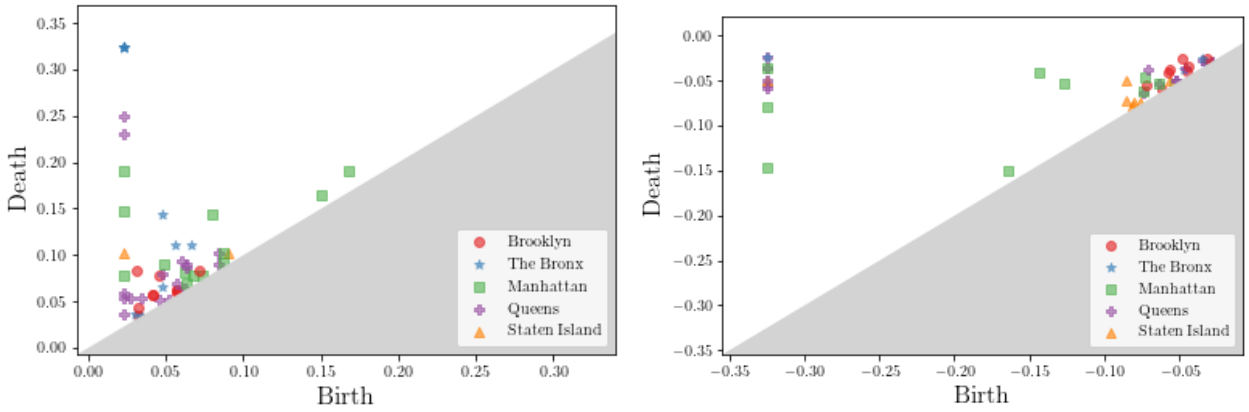
[16]At the time, the NYC Department of Health and Mental Hygiene defined "fully vaccinated" people to be individuals who either had received both doses of the Pfizer or Moderna vaccine or had received one dose of the Johnson & Johnson vaccine. (This differs from common parlance at that time, in which people were sometimes considered to be "fully vaccinated" only after two weeks had passed since their final dose of a vaccine.)

on 23 February 2021.

We do not possess the daily vaccination-rate data that is necessary to compute a vineyard, so instead we calculate the PH of $\mathcal{K}$ with the sublevel-set and superlevel-set filtrations from Sections 5.3.1 and 5.3.2. We show the resulting PDs for the 1D PH in Figure 5.4.2. As we described in Section 5.3.1, the points in the PD from the sublevel-set filtration correspond to zip codes in which vaccination rates are higher than in the neighboring zip codes. The death of a homology class is the vaccination rate in that zip code, and the birth of a homology class reflects the extent of spatial isolation of that zip code from other local maxima. An earlier birth filtration implies more spatial isolation. Similarly, the points in the superlevel-set filtration PD correspond to zip codes in which the vaccination rate is lower than in the neighboring areas. As we discussed in Section 5.3.1, we obtain the zip code that is associated with a homology class from its death simplex σ_d. We color the points in the PDs by the boroughs of their corresponding zip codes.

In Figures 5.4.3 and 5.4.4, we highlight the locations of the maxima and minima, respectively. In Figures 5.4.3a and 5.4.4a, we color the extrema based on their vaccination rates. In Figure 5.4.4a, we observe that the minima all have near-0 vaccination rates. In Figures 5.4.3b and 5.4.4b, we color each zip code according to the persistence (i.e., death $-$ birth) of its corresponding homology class. These two figures incorporate global information about the structure of the extrema, as we described in the paragraph above and in Section 5.3. For example, in Figure 5.4.4b, we observe that some of the minima (specifically, those with the largest persistence values) are significantly more spatially separated than others, even though all of the minima have similar vaccination rates. A larger persistence of a local minimum indicates a greater difference in the vaccination rate between the minimum and the neighboring zip codes. A zip code that is a local minimum with a larger persistence may have a greater inequity in vaccine access than its neighboring regions. Such insights may be useful for sociologists and policy makers.

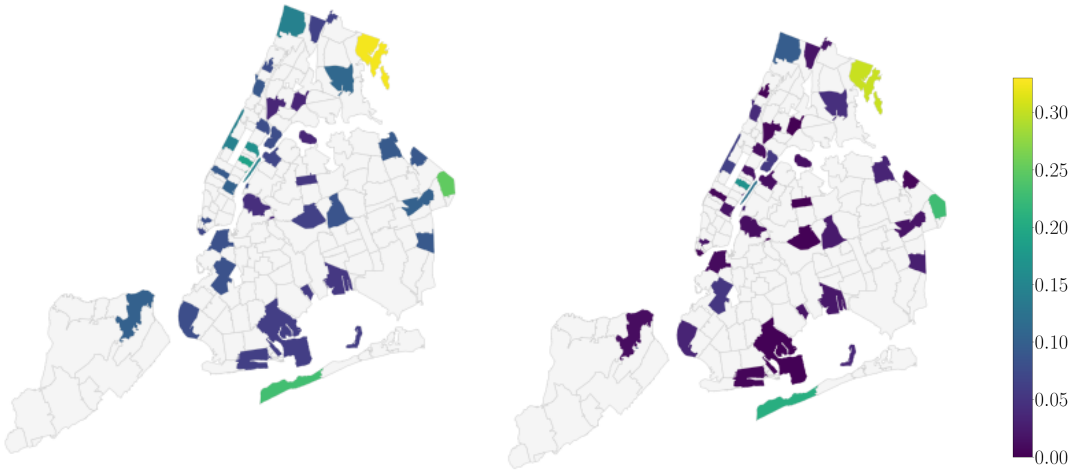
An issue arises from the fact that several of the NYC zip codes are islands and thus are



(a) Sublevel-set filtration

(b) Superlevel-set filtration

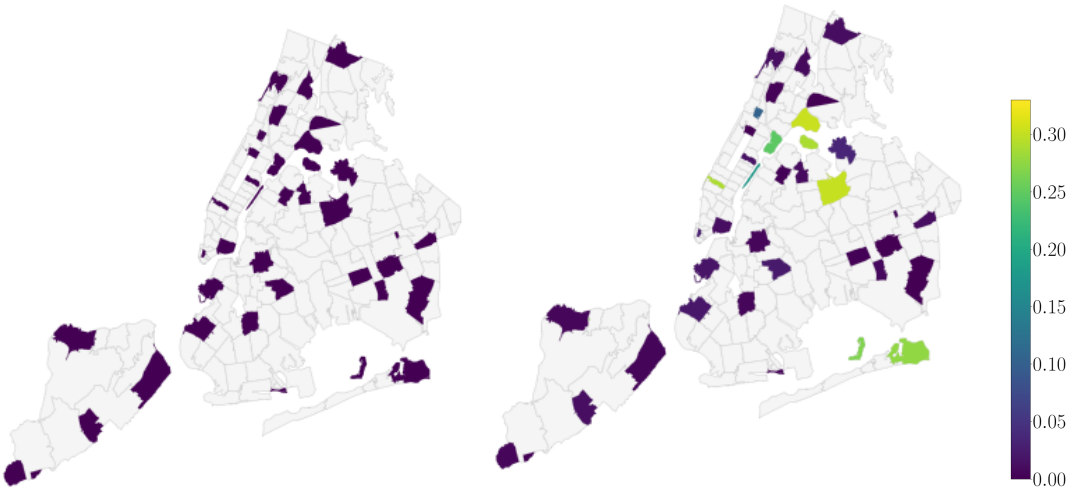
Figure 5.4.2: PDs for the 1D PH of the NYC simplicial complex with filtrations that are induced by the per capita full vaccination rate by zip code on 23 February 2021. We show only the finite homology classes. Each point in a PD corresponds to a zip code, which we label according to its borough [NYCa], that has (a) a higher vaccination rate than its neighboring zip codes or (b) a lower vaccination rate than its neighboring zip codes.



(a)

(b)

Figure 5.4.3: Maps of the local maxima of the NYC vaccination-rate function. (a) Color corresponds to the vaccination rate of a zip code. (b) Color corresponds to the persistence (i.e., death − birth) of the corresponding homology class.



(a) (b)

Figure 5.4.4: Maps of the local minima of the NYC vaccination-rate function. (a) Color corresponds to the vaccination rate of a zip code. (b) Color corresponds to the persistence (i.e., death − birth) of the corresponding homology class.

isolated. These islands are trivial extrema because they are not adjacent to any other zip codes. One may wish to exclude these trivial extrema from a PD. In appendix 5.B.2, we propose alternative methods for handling disconnected geographical spaces such as NYC.

One can use the PDs in Figure 5.4.2 to study inequities in vaccine access. For example, one may seek to discern patterns in demographic data that correspond to the most-persistent points in the PDs. For interested readers, we provide some demographic data in appendix 5.E.

5.4.2 COVID-19 Case Rate in the City of Los Angeles

We now examine time-dependent COVID-19 case rates in neighborhoods of the city of LA.[17] The geographical boundaries of the neighborhoods are given by a SHAPEFILE [Geo]. From

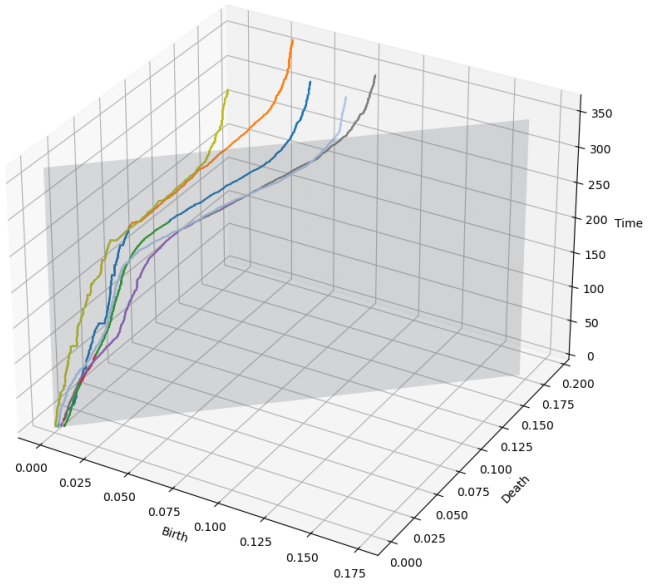
[17]We exclude Angeles National Forest because it has only 20 inhabitants.

the SHAPEFILE, we construct a simplicial complex $\mathcal{K}$ in the manner that we described in Section 5.2. We know the number of cases in each neighborhood on each day from 25 April 2020 to 25 April 2021. For each neighborhood, we divide the case count by the neighborhood population to obtain per capita case rates, and we calculate a running 14-day mean[18] on each day to smooth the data. For neighborhood R and time $t \in \{0, 1, \ldots, 365\}$, we define $F(t, R)$ to be the 14-day mean per capita case rate in R on day t after 25 April 2020. We compute the vineyard for a simplicial complex $\mathcal{K}$ using the time-dependent sublevel-set filtration that is induced by $F(t, R)$. We show the most important and interesting subsets of our vineyard in Figures 5.4.5 and 5.4.8. See Figure 5.C.1 for the full vineyard.

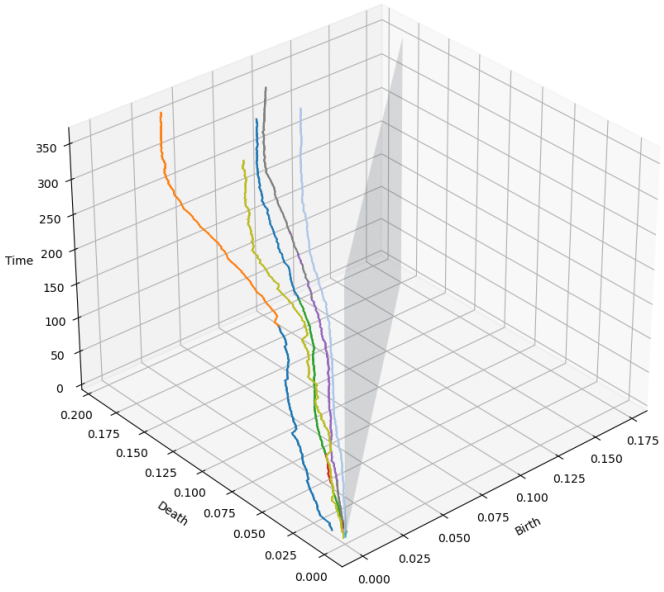
The vines in the vineyard correspond to COVID-19 anomalies, which we define to be neighborhoods that have a higher running 14-day mean COVID-19 case rate than the surrounding neighborhoods for at least one day. Anomalies that are more spatially isolated yield vines with earlier births, and anomalies with high case rates yield vines with late deaths. See Section 5.3.1 for a detailed discussion. We color each vine according to the geographical location(s) of its anomaly. As we discussed in Section 5.3.3, we obtain the anomaly location(s) from the time-dependent death simplex $\sigma_d(t)$ of a vine. The function $\sigma_d(t)$ is a piecewise-constant function; as it changes, so does the location of the associated anomaly. Therefore, the color of a vine can change with time. For example, consider Figure 5.4.5, where we show the five most-persistent vines.[19] The global maximum of the data set is initially in Little Armenia, but it moves to Vermont Square at about $t = 220$. In the vineyard, we see this from the vine that is initially blue (for Little Armenia) from time $t = 0$ until about $t = 220$ and then orange (for Vermont Square) starting from about time $t = 220$ through time $t = 365$. There are also other vines whose locations change with time. Such geographical location changes do not need to be adjacent, but they often are near each other. In Figure 5.4.7, we

[18]On day t, we take the mean of the case rates on days t, $t-1$, …, $t-13$. Some outlets (e.g., [STA]) report running 14-day means of COVID-19 case counts, and other outlets (e.g., [The]) report 14-day trends.

[19]Recall from Chapter 2.5 that the persistence of a vine is $\int_{t_0}^{T} [f(t, \sigma_d(t)) - f(t, \sigma_b(t))] \mathrm{d}t$.



(a)



(b)

Figure 5.4.5: (a) The five most-persistent vines of the vineyard for the LA simplicial complex with a sublevel-set filtration from the 14-day mean per capita case rate during the period 25 April 2020–25 April 2021. Each vine corresponds to a COVID-19 anomaly. We color each vine according to the geographical locations of its associated anomaly, which can change with time. (See Figure 5.4.6 for the legend.) (b) A different view of the same five vines.

Figure 5.4.6: The legend for Figure 5.4.5. Each of the depicted regions is a local maximum of the COVID-19 case-rate function for some subset of the time period 25 April 2020–25 April 2021.

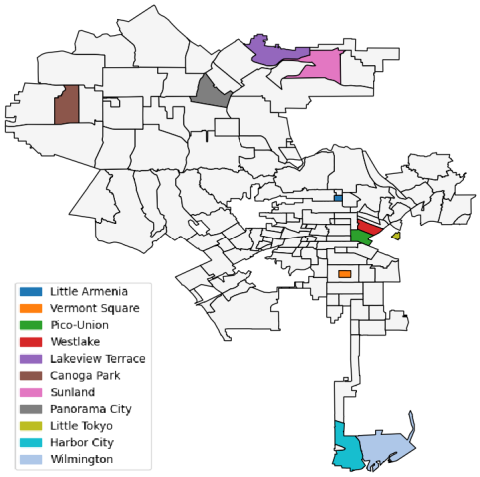


Figure 5.4.7: A map of the most-persistent anomalies of the COVID case-rate function in LA during the time period 25 April 2020–25 April 2021. Each of the highlighted regions is a local maximum of the COVID-19 case-rate function for some subset of the time period.

highlight these anomalies on a map.

A vineyard encodes the temporal persistence of anomalies. The length of time that a vine is not on the diagonal plane of a vineyard, which we henceforth call the "length" of a vine, is the amount of time that an anomaly exists in the vineyard. At the beginning of the COVID-19 pandemic, all neighborhoods had low per capita case rates. We expect an emerging anomaly to have a low case rate for a long time and then for the case rate to grow rapidly starting at some later time. An emerging anomaly in the "low-case-rate" phase yields a vine that is close to the diagonal for a long time. By examining the lengths of vines, we hypothesize that one can distinguish between concerning emerging anomalies (i.e., those

that may become major COVID-19 anomalies in the future) and anomalies of lesser concern, even when the anomalies have similar case rates.

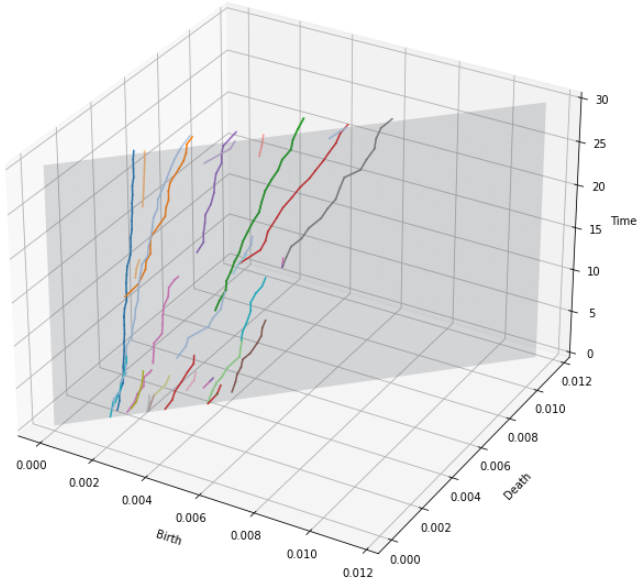
In Figure 5.4.8, we show case rates early in the time period that we track (and close to the "beginning"[20] of the COVID-19 pandemic) by computing the vineyard for the period 25 April 2020–25 May 2020. In the depicted vineyard, we exclude the 20 most-persistent vines to more easily see the vines that are close to the diagonal plane. Many of these latter vines are short, so their associated anomalies are short-lived. The longer vines are anomalies that are longer-lived and thus of greater concern in the long run, even though they are close to the diagonal during the period 25 April 2020–25 May 2020. For example, there is an anomaly at Wilmington that we show with the light-blue vine. This vine is close to the diagonal plane, but it has a large temporal persistence during the period 25 April 2020–25 May 2020. In Figure 5.4.5, we see that Wilmington eventually becomes one of the most-persistent anomalies in LA.

5.5 Discussion

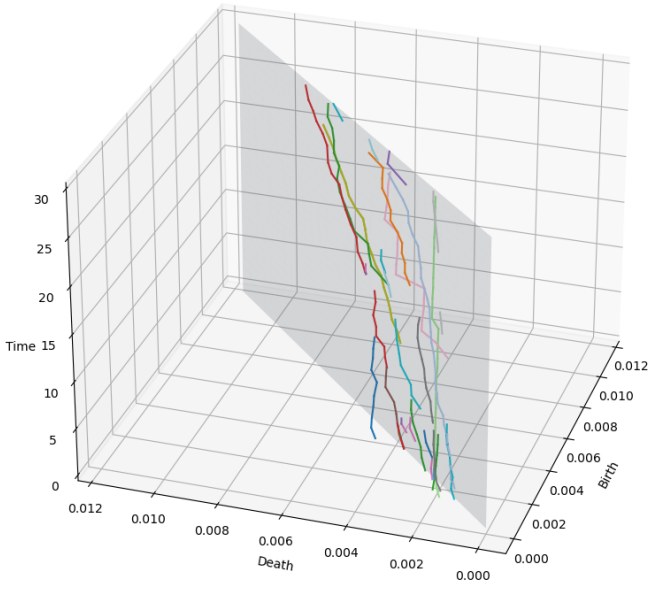
In our approach, we needed to make a variety of choices. There are other ways to construct a simplicial complex to represent a geographical space. There are also other choices of topological tools for analyzing time-varying data. We briefly discuss some of these possibilities in the next several paragraphs.

If one only cares about local information (specifically, the locations and values of the extrema) and not about global information (such as the spatial separation between extrema), then an alternative method for constructing a simplicial complex $\mathcal{K}$ is to construct the dual graph of the set S of regions. That is, for each region component R, there is a vertex $v_R \in \mathcal{K}_R$, and if regions R_1 and R_2 are adjacent, then there is an edge between v_{R_1} and v_{R_2}.

[20]The COVID-19 pandemic was declared a national emergency on 13 March 2020 [Tru], and the city of LA closed its public schools and ordered the closure of restaurants, bars, and gyms on 16 March 2020 [Kan].



(a)



(b)

Figure 5.4.8: (a) Vineyard for the LA simplicial complex with a sublevel-set filtration for the 14-day mean per capita case rate during the period 25 April 2020–25 May 2020. We exclude the 20 most-persistent vines to more easily see the vines that are near the diagonal plane. Each vine is associated with a COVID-19 anomaly, and we color each vine according to the geographical location(s) of its anomaly. See Figure 5.4.9 for the legend. (b) A different view.

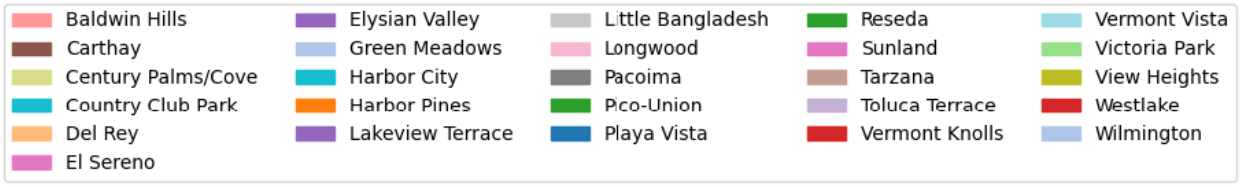


Figure 5.4.9: The legend for Figure 5.4.8. Each of the depicted regions is a local maximum of the COVID-19 case-rate function for some subset of the time period 25 April 2020–25 May 2020.

If we wish to study local maxima of a function $F : S \to \mathbb{R}$, then we define the filtration of an edge $e = (v_{R_1}, v_{R_2})$ to be $f(e) = \max\{F(R_1), F(R_2)\}$ and we define the filtration of a vertex v_R to be $f(v_R) = 0$. (There is an analogous definition for studying local minima.) In the 0D PH of the filtered complex $(\mathcal{K}, f)$, the homology classes correspond to local maxima. If a homology class's birth simplex is the vertex v_R, then R is the corresponding local maximum and $F(R)$ is the death of the homology class. All the 0D homology classes are born at 0; thus the birth does not provide any additional information, as it did for our construction in Section 5.2. Because of this, we do not obtain any global information from the PH of $(\mathcal{K}, f)$.

Rasterization gives another alternative method to construct a simplicial complex from SHAPEFILE data. When one rasterizes a SHAPEFILE, one can transform the resulting image into a simplicial complex by imposing the pixels of the image onto a triangulation of the plane. However, our approach has several key advantages over rasterization. First, the number of simplices in the simplicial complex that one obtains by rasterizing a SHAPEFILE is orders-of-magnitude larger than the number of simplices in our construction. Computing the PH of a simplicial complex with fewer simplices allows significantly faster computations. Second, the simplicial complex that one obtains by rasterization has no guarantee of "topological correctness," as property (P) may not hold. The extent to which the resulting simplicial complex is topologically correct depends on the resolution of the rasterization, and using a higher resolution requires more simplices. Our construction of simplicial complexes also

yields a natural way to map a 2D simplex to the geographical region that contains it. We use this preservation of geographical information to find the locations of the local extrema. Lastly, our construction allows us to detect anomalies on the boundary of a geographical space.

Our construction uses geographical adjacencies, but one may instead wish to employ "effective" distances between regions. One can calculate effective distances using mobility and transportation data. Two regions that are closely connected via transportation are effectively closer than they are based on direct geographical considerations; this affects phenomena such as the dynamics of infectious diseases [BH13, SBS21].

We used only 1D PH to study extrema, but one can alternatively use 0D PH if one is not interested in the geographical locations of the extrema; we discuss this in Appendix 5.B.1. In Appendix 5.B.2, we discuss alternative filtrations that one can apply to geographical spaces (such as NYC) that are disconnected. We used a time-dependent function on a geographical space to compute vineyards, but an alternative is to use an approach that is based on multiparameter PH. In Appendix 5.B.4, we discuss how one does this when the time-dependent function $F(\cdot, R)$ is monotonic for all regions R. When $F(\cdot, R)$ is not monotonic for all R, we discuss in Appendix 5.B.5 how one can use an approach that is based on multiparameter zigzag PH. Both multiparameter PH and multiparameter zigzag PH are difficult to visualize, and they both suffer from a lack of easily interpretable invariants. Consequently, we chose to compute vineyards for our applications.

5.6 Conclusions

We developed methods to directly incorporate spatial structure into applications of topological data analysis (specifically, of persistent homology) to geospatiotemporal and geospatial data. We defined a way to construct a simplicial complex that efficiently and accurately represents a geographical space. Given a function on a geographical space, we defined fil-

tration functions on a simplicial complex such that the homology classes are in one-to-one correspondence with either local minima or local maxima. By constructing a vineyard, one can track how the local extrema move and change with time.

We conducted case studies using COVID-19 vaccination and case-rate data. In one case study, we examined geospatial vaccination-rate structure in New York City on one day. In our other case study, in which we examined geospatiotemporal data, we constructed a vineyard to examine COVID-19 case-rate anomalies in the city of Los Angeles over the course of one year. From the vineyard, we identified the locations of these anomalies and measured the severity of the associated disease outbreaks. The vineyard also captures information about the relationships between anomalies, such as the extent to which they are isolated from each other. We calculated the temporal persistence of each anomaly based on the length of its corresponding vine.

There are several ways to build on our research. It is desirable to discover how to use a vineyard to produce systematic forecasts of how a disease (or something else) will spread in space and time. We hypothesized in Section 5.4.2 that one can identify "emerging anomalies" in the COVID-19 case-count data set as vines that are long but close to the diagonal plane. In other applications, one may wish to forecast which locations of local extrema will have the largest data values and/or the largest temporal persistences. One may also wish to forecast how extrema will move in space. It will be valuable to investigate how to use the output of our approach as an input to forecasting models.

Our approach is useful for a wide variety of applications, and it seems possible to generalize it for many others. For example, given spatiotemporal voting data, one can identify regions that vote differently from the neighboring regions. This would allow one to generalize the work of [FP21] to track the intensity of voting differences and study spatial relationships between different political islands. Our methodology is not restricted to geographical data. It is applicable whenever one has a surface that is partitioned into a finite number of regions and a real-valued function (or a sequence of real-valued functions) on those regions.

For example, it may be possible to apply our approach to grayscale image data by partitioning an image into regions in which pixel values are close to each other. It also seems possible to extend our approach to higher dimensions; this would require constructing a higher-dimensional simplicial complex when one has adjacency information for the higher-dimensional regions. For example, in three dimensions, one can use such an extension of our approach to study atmospheric, oceanic, and video dynamics.

APPENDIX

5.A Details of our Simplicial-Complex Construction

5.A.1 Boundary-Sequence Adjustment

Before constructing the polygons with holes P^R for each region R, we adjust the boundary sequences if necessary. The adjustment procedure proceeds as follows. Let $D_0^R, D_1^R, \ldots, D_{h_R}^R$ be the disks in the statement of assumption (A2), let $B_i^R = \partial D_i^R$, and let S_i^R denote the sequences of neighbors around B_i^R. First, we adjust the sequences so that, for each region R and each B_i^R, the first element of S_i^R has a 1D intersection with R. We then adjust the sequences so that $|S_i^R| \geq 3$ for all R and i. When $|S_i^R| < 3$, there are two cases:

1. (Case 1) If $|S_i^R| = 1$, let N be the unique element of S_i^R. This situation occurs if R is an island, and it can also occur if R lies inside N or if N lies inside R. We adjust S_i^R to be the sequence $\{N, N, N\}$. If N is not the exterior region, let j be the index such that B_j^N intersects R. Adjust S_j^N to be the sequence $\{R, R, R\}$ to compensate for the adjustment that we made to S_i^R.

2. (Case 2) If $|S_i^R| = 2$, let N_1 and N_2 be the two elements of S_i^R. If B_i^R intersects R, then R is adjacent to the exterior; without loss of generality, let N_1 denote the exterior region. For example, in Figure 5.2.3a, $S_0^{\text{Little Bangladesh}} = \{\text{Koreatown}, \text{Wilshire Center}\}$. We adjust S_i^R to be the sequence $\{N_1, N_1, N_2\}$. If N_1 is not the exterior region, which occurs if R is not adjacent to the exterior, then we also adjust $S_j^{N_1}$ to compensate, where j is the index of the boundary component of N_1 that intersects R. In this case, we adjust $S_j^{N_1}$ by repeating R an additional time.

5.A.2 Construction of $\mathcal{K}$ from the Set $\{P^R \mid R \in S\}$

We present two lemmas that we used in Section 5.2 to construct $\mathcal{K}$ by gluing together the set $\{P^R \mid R \in S\}$ of polygons with holes.

Lemma 5.A.1. Let R_1 and R_2 be connected regions in a set S that satisfies assumptions (A1)–(A4). Let $D_0, \ldots, D_h$ be the disks in the statement of (A2) for R_1. It is then the case that exactly one of the following statements is true:

1. $R_2 \subseteq \text{int}(D_0)^c$ and $R_2 \cap \text{int}(D_i) = \emptyset$ for all $i > 0$; or
2. there is an $i > 0$ such that R_2 is enclosed in D_i and $R_2 \cap \text{int}(D_j) = \emptyset$ for all $j \neq i$.

Proof. Because the interiors of R_1 and R_2 do not intersect, it must be true that $\text{int}(R_2) \subseteq \text{int}(D_0)^c \cup \left(\bigcup_{i=1}^{h} \text{int}(D_i) \right)$. Therefore,

$$\text{int}(R_2) = \left(\text{int}(D_0)^c \cap \text{int}(R_2) \right) \cup \left(\bigcup_{i=1}^{h} \text{int}(D_i) \cap \text{int}(R_2) \right).$$

The claim follows because $\text{int}(R_2)$ is connected and $\text{int}(D_0)^c, \text{int}(D_1), \ldots, \text{int}(D_h)$ are pairwise disjoint. □

Lemma 5.A.2. Let P^R be the annotated polygon with holes for a connected region R, let v be a vertex in P^R, and let $\{R, N_1, \ldots, N_n\}$ be the sequence of region adjacencies for v. If $n \geq 2$ and $N_1, \ldots, N_n$ are connected, then P^R has at most one other vertex w with the same set of region adjacencies. Additionally, if w exists, its sequence of region adjacencies must be $\{R, N_n, \ldots, N_1\}$, which is the mirror of the orientation of neighbors around v.

Proof. Suppose that w is another vertex in P^R with the same set of region adjacencies as v. Let v' and w' denote the points on the boundary of R that correspond, respectively, to v and w. Let R_0 be any connected region that is adjacent to both v' and w', let $D_0, D_1, \ldots, D_h$ denote the disks in the statement of (A2) for R_0, and let $B_i = \partial D_i$. Suppose that v' is in B_i. If $i = 0$, then there is a neighboring region N that is contained entirely in $\text{int}(D_0)^c$ (by

Lemma 5.A.1) and adjacent to v'. If $i > 0$, then there is a neighboring region N that is contained entirely in $\text{int}(D_i)$ (by Lemma 5.A.1) and adjacent to v'. In either case, $w' \in B_i$ because w' is also adjacent to N. Let $B_{i_1}, \ldots, B_{i_m}$ be the boundaries that contain v'. As we just showed, it must also be true that $w' \in B_{i_1}, \ldots, B_{i_m}$. If $m > 1$, then $w' \notin B_{i_1} \cap \cdots \cap B_{i_m}$ because $D_{i_1} \cap \cdots \cap D_{i_m}$ is a single point by assumption (A2); this is a contradiction. This argument shows that if v and w have the same set of region adjacencies, then there is a unique B_i that contains v', there is a unique B_j that contains w, and $B_i = B_j$.

Let B be the disk boundary of R that contains v and w. Either the interior of R is contained in the region that is bounded by B or it is contained in the complement of the region that is bounded by B. Without loss of generality, we suppose that the former is true. Let π be the permutation of $\{1, \ldots, n\}$ such that the sequence of region adjacencies around w is $\{R, N_{\pi(1)}, \ldots, N_{\pi(n)}\}$. Let $i_1, i_2 \in \{1, \ldots, n\}$, with $i_1 < i_2$, be a pair of indices. By the argument above (with $R_0 = N_{i_1}$), there is a unique disk boundary B_1 for N_{i_1} that contains v' and w'. Similarly, there is a unique disk boundary B_2 for N_{i_2} that contains v' and w'. We have that $v', w' \in B_1 \cap B_2$.

Because B_1 is homeomorphic to S^1, there exist paths γ_1 and γ_2 from v' to w' such that $\gamma_1 \cup \gamma_2 = B_1$. Because the interior of N_{i_1} does not intersect R, it follows that γ_1 and γ_2 are both in the complement of the region that is bounded by B'. There are two paths from v' to w' on B'. Let τ be the unique choice of path such that R is not contained in the region that is bounded by the closed curve $\tau \cup \gamma_1$. Either γ_1 is in the region that is bounded by the closed curve $\tau \cup \gamma_2$ or γ_2 is in the region that is bounded by the closed curve $\tau \cup \gamma_1$. Without loss of generality, we suppose that the latter is true.

Analogously to our argument above, there exist paths γ_3, γ_4 from v' to w' such that $\gamma_3 \cup \gamma_4 = B_2$ and γ_3 and γ_4 are in the complement of the region that is bounded by B. Because B_2 is homeomorphic to S^1, the paths γ_3 and γ_4 are either both contained in the region that is bounded by $\gamma_1 \cup \tau$ or both contained in the complement of the region that is bounded by $\gamma_2 \cup \tau$. Because $i_2 > i_1$, it must be the former case. Therefore, $\pi(i_2) < \pi(i_1)$.

It follows that π is order-reversing. If there were another vertex x in B that is adjacent to the same set of regions, then the orientation of those regions around x would be the mirror of both the orientation of regions around v and the orientation of regions around w. This gives a contradiction when $n \geq 2$. □

For example, let R be the region Koreatown in Figure 5.2.3a. The two vertices that are shared by Koreatown and Little Bangladesh have the same region adjacencies, but they have mirrored orientations.

5.B Alternative Topological Approaches

5.B.1 0D Persistent Homology

We do not compute 0D PH in the present chapter. However, it is appropriate to use 0D PH to study the structure of local extrema when one is not interested in their geographical locations.

Let F be a real-valued function on a set S of geographical regions. In Section 5.3.1 (respectively, Section 5.3.2), we described how one can analyze the local maxima (respectively, local minima) of F by computing the 1D PH of the sublevel-set filtration (respectively, superlevel-set filtration). We now discuss how the 0D PH of the sublevel-set filtration (respectively, superlevel-set filtration) yields information about local minima (respectively, local maxima) of F.

The 0D PH of the sublevel-set filtration encodes information about the structure of local minima of F in a way that is similar to how 1D PH encodes information about the structure of local maxima. One can imagine taking α-sublevel sets of the function in Figure 5.3.2 (where we display α-*super*level sets) to see why this is true. A region R is a *local minimum* if $F(R)$ is less than $F(N)$ for all neighboring regions N of R for which $N \cap R$ is 1D. If R is a local minimum, there is a 0D homology class whose birth simplex is one of the vertices in

one of the triangles in the preimage $g^{-1}(R)$. The class is born at filtration parameter-value $\alpha = F(R)$. For the LA data set of COVID-19 case rates, 0D homology classes correspond to regions that have a lower case rate than neighboring regions. The smaller the value $F(R)$ in comparison to the neighboring regions, the more persistent the homology class is. There is also one infinite 0D homology class for each connected component. One can think of these classes as corresponding to a "local minimum" in the exterior region. However, unlike for 1D homology classes, there is no canonical map from 0D homology classes to regions because the birth simplex of a 0D class is a vertex that belongs to several regions. Analogously, the 0D PH of the superlevel-set filtration encodes information about the structure of local maxima of F. However, as with a sublevel-set filtration, there is no canonical map from 0D homology classes to regions. Therefore, one cannot easily use the 0D PH of the sublevel-set filtration (respectively, superlevel-set filtration) to identify the geographical locations of the local minima (respectively, local maxima), so we did not examine 0D PH in our case studies.

5.B.2 Alternative Filtrations for Disconnected Geographical Spaces

In Section 5.3.1 (respectively, Section 5.3.2), we defined a sublevel-set filtration (respectively, superlevel-set filtration) in which we set the filtration values of all exterior-adjacent vertices and edges to the global minimum (respectively, to the additive inverse of the global maximum) of F. In applications in which the union of all regions is not connected, such as for the NYC zip codes in Section 5.4.1, an alternative definition is to consider extrema on each connected component separately, rather than on the entire geographical space at once. This solves the problem that an isolated region (i.e., a geographical island[21]) is trivially both a local maximum and a local minimum because it is not adjacent to any other regions. In Definitions 5.3.1 and 5.3.2, they appear as 1D homology classes that are born at the earliest filtration time; this may falsely emphasize the persistence of these trivial extrema.

[21]These are literal islands, rather than "islands" from a PH computation.

Definition 5.B.1 (Alternative Sublevel-Set Filtration)**.** Let $\mathcal{K}$ be the simplicial complex from Section 5.2 for a set S of regions, and let g be the assignment of 2D simplices to regions. Additionally, let $F : S \to \mathbb{R}$. If σ is a vertex or edge on the boundary of $\mathcal{K}$, let $\tilde{\sigma}$ be the 2D simplex for which σ is on the boundary of $\tilde{\sigma}$. On σ, we define the alternative sublevel-set filtration function f to be

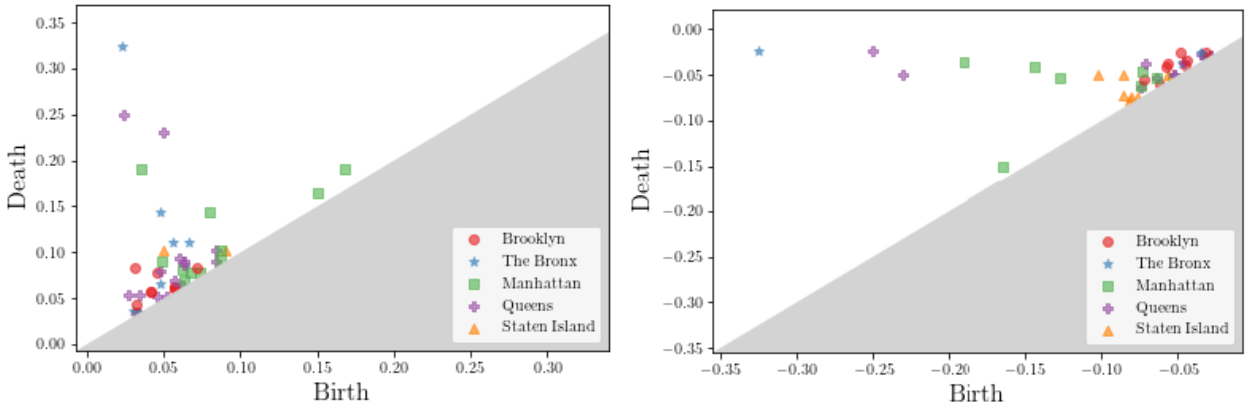
$$f(\sigma) = \min_{R} \{F(R) \mid R \subseteq C\},$$

where C is the connected component that contains the region $g(\tilde{\sigma})$. On all other simplices, the filtration function f is equal to the sublevel-set filtration function.

Definition 5.B.2 (Alternative Superlevel-Set Filtration)**.** Let $F : S \to \mathbb{R}$ for a set S of regions. The alternative superlevel-set filtration function f is the alternative sublevel-set filtration function that is induced by $-F$.

Definitions 5.B.1 and 5.B.2 are appropriate options if one seeks to treat each connected component independently. In these alternative definitions, each connected component uses only information about other regions in the same component. One then compares region values $F(R)$ to global extremum values on their connected components. One consequence of using these definitions is that one ignores isolated regions, which are trivial extrema. In Definitions 5.B.1 and 5.B.2, these isolated extrema appear as points on the diagonal of a PD. This is often an appropriate way to handle isolated regions. However, when an isolated region is a global extremum of a data set, this may be undesirable. This situation never occurs in our data.

NYC has 14 connected components; several of them are zip codes that correspond to isolated islands. The alternative sublevel-set and superlevel-set filtrations effectively treat each connected component of NYC separately. In Figures 5.B.1a and 5.B.1b, we show the PDs that we compute using the alternative sublevel-set and superlevel-set filtrations that are induced by the vaccination-rate function that we defined in Section 5.4.1. In these PDs, we compare a zip code's per capita vaccination rate to the global minimum or maximum



(a) Alternative sublevel-set filtration

(b) Alternative superlevel-set filtration

Figure 5.B.1: PDs for the 1D PH of the NYC simplicial complex with filtrations that are induced by the per capita full vaccination rate by zip code on 23 February 2021. We show only the finite homology classes. Each point in a PD corresponds to a non-isolated zip code, which we label according to its borough [NYCa], that has (a) a higher vaccination rate than its neighboring zip codes or (b) a lower vaccination rate than its neighboring zip codes.

rate on its connected component, rather than to the global extremum in all of NYC. More precisely, the birth time of a connected component's global extremum is either the lowest per capita vaccination rate of that component (for the alternative sublevel-set filtration) or the additive inverse of the highest per capita vaccination rate of that component (for the alternative superlevel-set filtration). Consequently, the trivial island extrema yield homology classes on the diagonal of a PD.

The alternative sublevel-set filtration and the alternative superlevel-set filtration, along with time-dependent versions of them, are implemented in our code at `https://bitbucket.org/ahickok/vineyard/src/main/`.

5.B.3 Level-set complexes

This appendix is adapted from [FHP22] and consists of only the contributions that I led.

In this appendix, we use a level-set construction of simplicial complexes, which were introduced in [FP21], to study anomalies in COVID-19 infections.

5.B.3.1 Definition of a Level-Set Complex

We now briefly review the level-set construction of filtered complexes that was introduced in [FP21]. In a level-set filtration, one describes data as a manifold. Let M denote a two-dimensional (2D) manifold, such as data in an image format. We construct a sequence

$$M_0 \subseteq M_1 \subseteq \cdots \subseteq M_n$$

of manifolds (where M_0 is an approximation of M) as follows. At each time t, we evolve the boundary Γ_t of M_t outward according to the level-set equation of front propagation [OF03]. Specifically, for a manifold M that is embedded in $\mathbb{R}^2$, we define a function $\phi(\vec{x}, t) \colon \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$, where $\phi(\vec{x}, t)$ is the signed distance function from $\vec{x}$ to Γ_t at time $t \geq 0$. We propagate Γ_t outward at velocity v using the partial differential equation

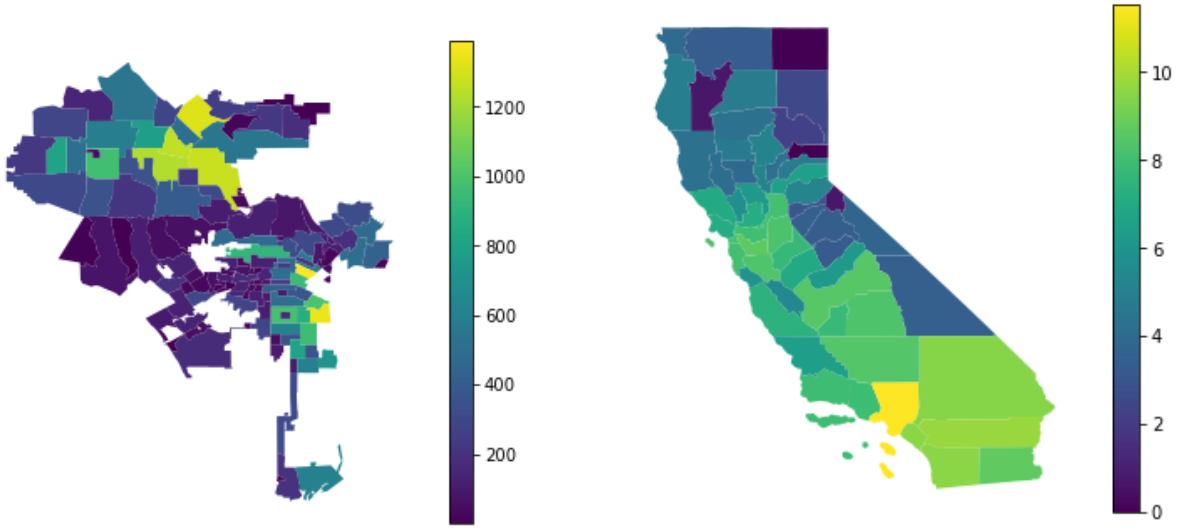
$$\frac{\partial \phi}{\partial t} = v|\nabla \phi| \tag{5.2}$$

until all homological features die. The evolution (5.2) gives a signed distance function at each time t. We take M_t to be the set of points $\vec{x}$ such that $\phi(\vec{x}, t) > 0$. (This corresponds to points inside the boundary Γ_t.) In our examples in this appendix, we use $v = 1$.

By imposing $\{M_i\}$ over a triangular grid of points (see [FP21]), we obtain a corresponding simplicial complex $\mathcal{K}_i$ for each M_i. Because the level-set equation (5.2) evolves outward, we satisfy the condition that $\mathcal{K}_i \subseteq \mathcal{K}_{i+1}$ for all i, so $\{\mathcal{K}_i\}$ is a filtered complex.

5.B.3.2 An Application of Level-Set Complexes to COVID-19 Data

We use PH to analyze the spatial properties of the spread of COVID-19 in Los Angeles (LA) neighborhoods and California counties. In contrast to the main part of this chapter, we do not consider spatiotemporal dynamics. We consider two data sets. The first is a highly



(a) LA Neighborhoods

(b) California Counties

Figure 5.B.2: Cumulative COVID-19 case counts on 30 June 2020 in (a) Los Angeles neighborhoods and (b) California counties. We plot the LA case counts on a linear scale and the California county case counts on a (natural) logarithmic scale.

granular data set that consists of COVID-19 case counts in 136 LA neighborhoods on 30 June 2020. The second is a coarser data set that consists of case counts in the 58 counties of California on the same day [Fac]. For each data set, we also have geographic information in the form of a SHAPEFILE [Por19, Geo]. We visualize these data sets in Figure 5.B.2.

Let M^{LA} denote the 2D manifold that consists of the union of LA neighborhoods with fewer than 750 cumulative cases, and let M^{CA} denote the union of California counties with fewer than 5,000 cumulative cases. We approximate these manifolds by rasterizing the associated SHAPEFILES to obtain manifolds M_0^{LA} and M_0^{CA}. We show M_0^{LA} and M_0^{CA} in Figure 5.B.3. As we described in Section 5.B.3.1, we construct sequences of manifolds starting from M_0^{LA} and M_0^{CA} using level-set dynamics (5.2). We then construct a level-set filtration for each of these sequences by imposing the manifolds in them on a triangulation of the plane.

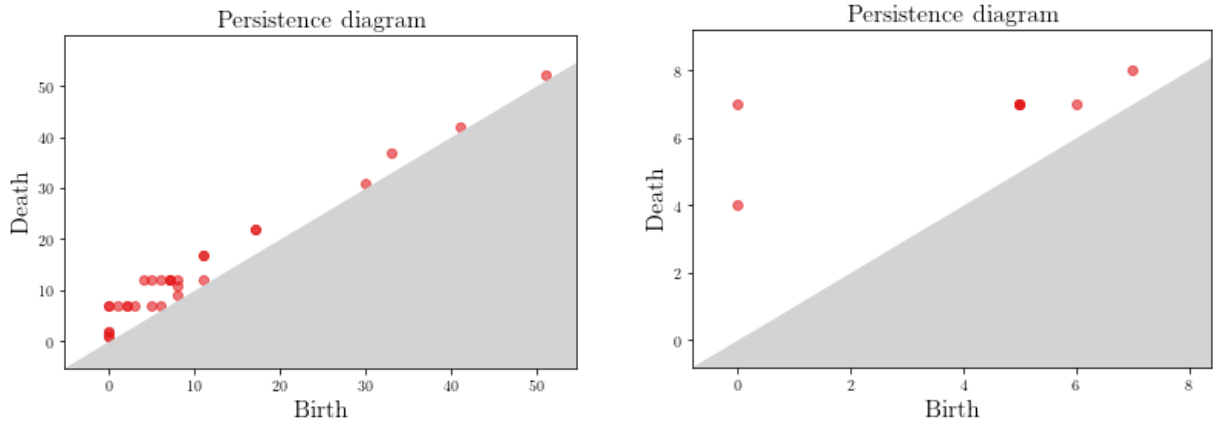
(a) M_0^{LA} (b) M_0^{CA}

Figure 5.B.3: Initial manifolds for the level-set filtrations that we construct from data of the spread of COVID-19. (a) The manifold M_0^{LA} is an approximation of the manifold M^{LA}, which consists of the union of LA neighborhoods with fewer than 750 cumulative cases on 30 June 2020. (b) The manifold M_0^{CA} is an approximation of the surface M^{CA}, which consists of the union of California counties with fewer than 5,000 cumulative cases on 30 June 2020.

In Figure 5.B.4, we show the PDs that we compute for the 1D PH of our level-set complexes for the two data sets. These PDs can help us identify COVID-19 anomalies. We define an "anomaly" to be a set of geographical regions (neighborhoods in the LA data and counties in the California data) in which the case count is higher than in the surrounding area. This notion of an anomaly is analogous to the political "islands" that were studied using PH in [FP21]. An anomaly with a case count that is at least as large as the threshold (750 for LA neighborhoods and 5,000 for California counties) appears as a 1D hole in M_0, unless the anomaly is adjacent to the boundary of the map (e.g., LA County in California). The anomalies that are not adjacent to the boundary correspond to homology classes that are born at time 0. The persistence of the corresponding 1D homology class is proportional to the geographic size of the anomaly. Therefore, the PDs reflect both the number of anomalies and the sizes of the anomalies.

There is not a one-to-one correspondence between anomalies and homology classes. Homology classes that are born after time 0 usually reflect only the geography of the regions, although they sometimes correspond to anomalies on the boundary of the map. There is also not a one-to-one correspondence between anomalies and homology classes that are born at time 0. Some of the homology classes that are born at 0 are simply holes in the map (e.g., see Figure 5.B.2a), and anomalies that are adjacent to a boundary do not necessarily correspond to any homology class.

The analysis in this appendix shows that one can use a level-set filtration to study the number and sizes of COVID-19 anomalies on both a granular level (by considering neighborhoods in Los Angeles) and a coarse level (by considering counties in California). We used only case counts in our computations, but one can also construct level-set filtrations for the death counts, hospitalization counts, or other quantities. The level-set filtration is flexible, but this approach has important limitations. For example, we only detected anomalies with a case count that is above some fixed threshold. This restricts us to measuring the severity of an outbreak based on its geographic area. One way to address this issue is



(a) LA neighborhoods

(b) CA counties

Figure 5.B.4: The PDs for the 1D PHs of the level-set filtrations for COVID-19 cases in (a) LA and (b) California.

by applying the level-set filtration after constructing a cartogram [GN04], instead of directly from a SHAPEFILE. Additionally, the level-set filtration is unable to detect anomalies that occur on the boundary of a map (e.g., Los Angeles County when considering counties in California). These limitations were addressed by the methods in the main part of this chapter.

5.B.4 Multiparameter Persistent Homology

One can use multiparameter persistent homology (MPH) to study how the topology of a data set changes as one varies multiple parameters. For a review of MPH, see [CZ07, BL22].

One can use MPH to study local extrema of functions that are nondecreasing with time. To apply MPH to our COVID-19 case-rate data, two feasible parameters are (1) time and (2) the cumulative COVID-19 case rate. However, MPH is difficult to analyze. Although there are invariants (e.g., the rank invariant), there is no complete discrete invariant [CZ07]. By contrast, one can use PDs for single-parameter PH.

Definition 5.B.3. Let $\mathcal{K}$ be the simplicial complex from the construction in Section 5.2 for a set S of regions. Let $F : \{t_0, \ldots, t_n\} \times S \to \mathbb{R}$ be a function such that $F(t, R) \geq F(s, R)$ for all $t \geq s$. Define the function $f(t_i, \sigma)$ to be the sublevel-set filtration that is induced by $F(t_i, \cdot)$. Let $\{\alpha_0, \ldots, \alpha_\ell\}$ be the image of F, where $\ell + 1$ is the number of elements in the image. We define the bifiltration

$$\mathcal{K}_{i,j} := \begin{cases} \{\sigma \in \mathcal{K} \mid f(t_i, \sigma) \leq \alpha_j\}\,, & i \in \{0, \ldots, n\}\,,\ j \in \{0, \ldots, \ell\} \\ \mathcal{K}\,, & j > \ell \text{ and } i \geq 0 \\ \mathcal{K}_{n,j}\,, & i > n \text{ and } j \geq 0 \\ \emptyset\,, & i < 0 \text{ or } j < 0\,. \end{cases}$$

One can use Definition 5.B.3 to study cumulative COVID-19 case rates as a function of time.

5.B.5 Multiparameter Zigzag Persistent Homology

One can use *multiparameter zigzag PH* (MZPH) to study how the topology of a data set changes as one varies multiple parameters nonmonotonically. See Section 2.1 of [CS10] for a short discussion of MZPH.

To use MZPH to study our COVID-19 case-rate data, two feasible parameters are (1) time and (2) the current COVID-19 case rate. A diagram of simplicial complexes, such as in Equation 5.3, induces a diagram of homology groups. This is a representation of a quiver. However, there are no known well-behaved statistical summaries (in contrast to single-parameter zigzag PH).

Definition 5.B.4. Let $\mathcal{K}$ be the simplicial complex from the construction in Section 5.2 for a set S of regions, and suppose that $F : \{t_0, \ldots, t_n\} \times S \to \mathbb{R}$. Define half steps $t_{i+1/2} := t_i + (t_{i+1} - t_i)/2$ for $i \in \{0, \ldots, m-1\}$, and let $s_i := t_{i/2}$. Define the function

$G : \{s_0, \ldots, s_{2n}\} \times S \to \mathbb{R}$ as follows:

$$G(s_i, R) = \begin{cases} F(s_i, R)\,, & i \text{ is even} \\ \max\{F(t_{(i-1)/2}, R), F(t_{(i+1)/2}, R)\}\,, & i \text{ is odd}\,. \end{cases}$$

We define the function $h(s_i, \cdot)$ to be the sublevel-set filtration that is induced by $G(s_i, \cdot)$. Let $\{\alpha_0, \ldots, \alpha_\ell\}$ be the image of G. We define

$$\mathcal{K}_{i,j} := \begin{cases} \{\sigma \in \mathcal{K} \mid h(s_i, \sigma) \leq \alpha_j\}\,, & i \in \{0, \ldots, 2n\}\,,\, j \in \{0, \ldots, \ell\} \\ \mathcal{K}\,, & j > \ell \text{ and } i \geq 0 \\ \mathcal{K}_{2n,j}\,, & i > 2n \text{ and } j \geq 0 \\ \emptyset\,, & i < 0 \text{ or } j < 0\,. \end{cases}$$

This yields the following diagram:

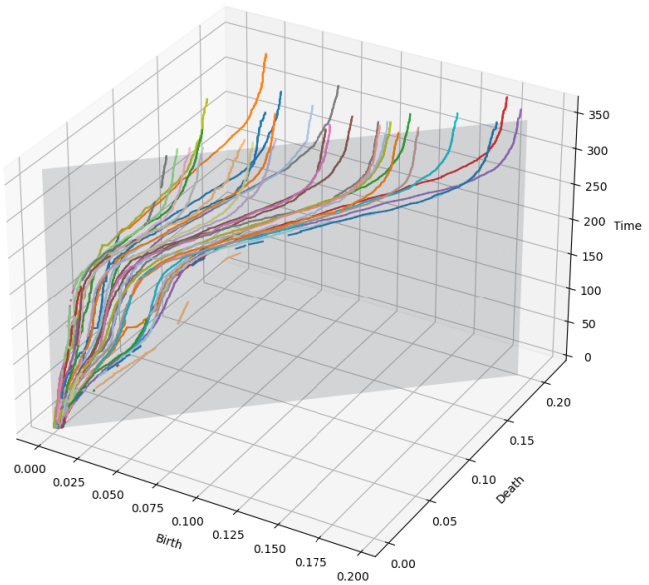
$$\begin{array}{ccccccc} \uparrow & & \uparrow & & \uparrow & & \uparrow \\ K_{\alpha_0,s_3} & \hookrightarrow & K_{\alpha_1,s_3} & \hookrightarrow & K_{\alpha_2,s_3} & \hookrightarrow & K_{\alpha_3,s_3} & \hookrightarrow \\ \downarrow & & \downarrow & & \downarrow & & \downarrow \\ K_{\alpha_0,s_2} & \hookrightarrow & K_{\alpha_1,s_2} & \hookrightarrow & K_{\alpha_2,s_2} & \hookrightarrow & K_{\alpha_3,s_2} & \hookrightarrow \\ \uparrow & & \uparrow & & \uparrow & & \uparrow \\ K_{\alpha_0,s_1} & \hookrightarrow & K_{\alpha_1,s_1} & \hookrightarrow & K_{\alpha_2,s_1} & \hookrightarrow & K_{\alpha_3,s_1} & \hookrightarrow \\ \downarrow & & \downarrow & & \downarrow & & \downarrow \\ K_{\alpha_0,s_0} & \hookrightarrow & K_{\alpha_1,s_0} & \hookrightarrow & K_{\alpha_2,s_0} & \hookrightarrow & K_{\alpha_3,s_0} & \hookrightarrow \end{array} \quad . \tag{5.3}$$

The inclusion maps induce a corresponding diagram of homology groups.

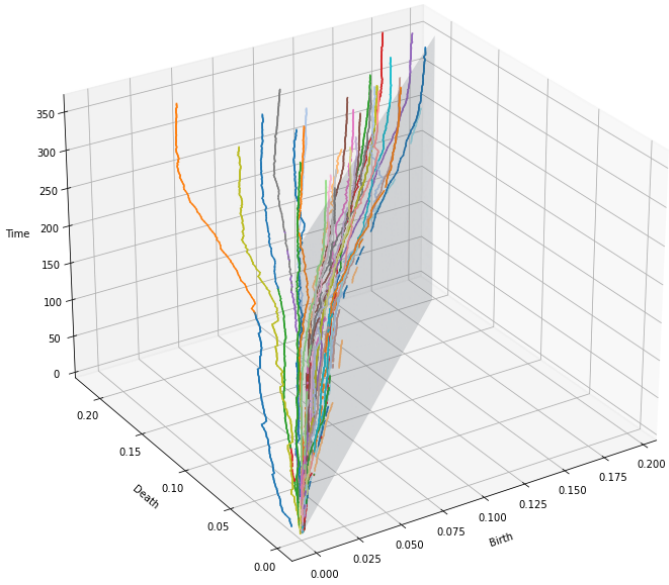
One can use Definition 5.B.4 to study non-cumulative COVID-19 case rates as a function of time.

5.C The Full LA Vineyard

In Figure 5.C.1, we show the full vineyard that we discussed in Section 5.4.2.



(a)



(b)

Figure 5.C.1: (a) The vineyard for the LA simplicial complex that we construct using the sublevel-set filtration from the 14-day mean per capita case rate during the period 25 April 2020–25 April 2021. Each vine is associated with a COVID-19 anomaly. We color each vine according to the geographical location(s) of its associated anomaly, which can change with with time. (See Figure 5.C.2 for the legend.) (b) A different view of the same vineyard.

Figure 5.C.2: The legend for Figure 5.C.1. Each of the depicted regions is a local maximum of the COVID-19 case-rate function for some subset of the time period 25 April 2020–25 April 2021.

5.D Results of an All-But-One Statistical Test

In Chapter 5, we examined local extrema of real-valued geospatial data; we called these "anomalies". For real-valued geospatiotemporal data, one can alternatively examine a different notion of anomaly. In this context, we say that a region is an *anomaly* if one is not able to infer its data successfully from the data of the other regions. More precisely, let X be the matrix whose (i, j)th entry is the value of region j at time step i. In our case study of COVID-19 case rates in LA, the regions are the neighborhoods of LA and the (i, j)th entry of X is the 14-day mean per capita case rate in region j on the ith day after 25 April 2020. Let $\boldsymbol{x}^j$ denote the jth column of X, and let X^j denote the matrix that one obtains by deleting column $\boldsymbol{x}^j$. The vector $\boldsymbol{x}^j$ has the data for region j, and the matrix X^j has the data for all regions except for region j. We define our *prediction* of region j to be the least-squares solution $\boldsymbol{b}^*$ to $X^j\boldsymbol{b} = \boldsymbol{x}^j$, and we quantify the predictability of region j by calculating the relative residual norm $\left\|X^j\boldsymbol{b}^* - \boldsymbol{x}^j\right\|_2 / \left\|\boldsymbol{x}^j\right\|_2$. A smaller relative residual norm

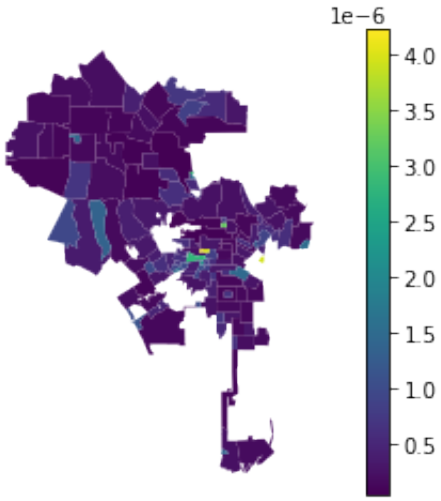


Figure 5.D.1: The results of an all-but-one statistical test for the LA COVID-19 case-rate data. We plot the relative residual norm for each neighborhood.

indicates greater predictability.

In Figure 5.D.1, we show the result of this "all-but-one" statistical test for the LA COVID-19 data set. In this figure, we plot the relative residual norm for each neighborhood. All neighborhoods have near-0 relative residual norms, so the neighborhoods' case rates are very predictable when one knows the case rates of all other neighborhoods. The mean relative residual norm is only 5.970×10^{-7}, with a standard deviation of $\sigma \approx 7.558 \times 10^{-7}$. The neighborhoods with the least predictability (specifically, those whose relative residual norms have a z-score that is larger than 3) are Brookside, Little Armenia, Little Tokyo, Sycamore Square, and Toluca Terrace. We show their relative residual norms and z-scores in Table 5.D.1.

The difference between what we learn from the all-but-one statistical test and what we learn from our TDA approach is the following. Using our TDA approach, we identified local extrema (i.e., regions whose values are either all larger than or all smaller than those of all neighboring regions); this is a geographical notion of anomaly. By contrast, the all-but-one statistical test does not inherently capture local extrema because the test does not con-

Neighborhood	Relative Residual Norm	z-score
Brookside	3.973×10^{-6}	4.466
Little Armenia	3.220×10^{-6}	3.471
Little Tokyo	3.944×10^{-6}	4.429
Sycamore Square	3.944×10^{-6}	4.429
Toluca Terrace	2.873×10^{-6}	3.012

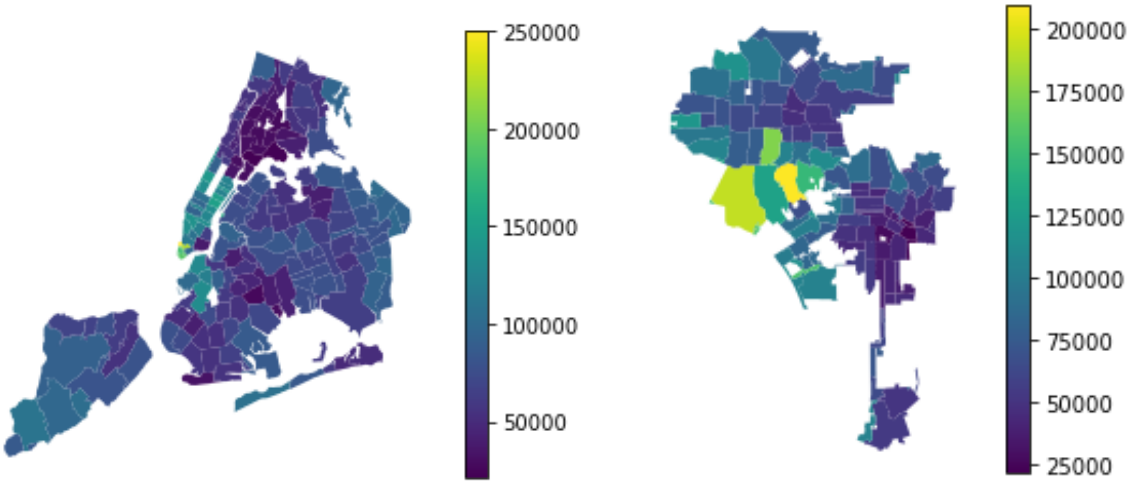
Table 5.D.1: The relative residual norms and z-scores for the LA neighborhoods that are least predictability according to our all-but-one test.

sider geographical adjacencies. Despite this conceptual difference, we observe some overlap between the anomalies that the two approaches identify. For example, the neighborhoods Little Tokyo and Little Armenia are identified as anomalies by both approaches. For further examples, compare Figure 5.D.1 with Figure 5.4.7.

5.E Demographic Data

We provide some demographic data for NYC and LA for readers who are interested in studying patterns between the PDs and demographic data, although an investigation of such patterns is beyond the scope of the present study. In Figure 5.E.1, we plot the median income for each zip code[22] [US19]. The geographical boundaries of the NYC and LA zip codes are given by the SHAPEFILES [NYCc] and [Cor], respectively. It is worthwhile to examine and compare other demographic data (such as racial, religious, and political data) to the PDs.

[22]We do not possess median income data for LA zip codes 90073, 90089, 90095, 91330, 91522, and 91608. These zip codes are in non-residential areas. For example, 90073 corresponds to the Veterans Administration.



(a) (b)

Figure 5.E.1: (a) Median household income by zip code in NYC. (b) Median household income by zip code in LA.

Acknowledgments

We thank Henry Adams, Heather Zinn Brooks, Michelle Feng, Lara Kassab, and Nina Otter for helpful discussions. Additionally, we are grateful to Michelle Feng for teaching us how to work with geospatial data. We thank the Los Angeles County Department of Public Health for providing the LA city data on COVID-19 and the population estimates of LA neighborhoods.

CHAPTER 6

Persistent Homology for Resource Coverage: A Case Study of Access to Polling Sites

This chapter is adapted from [HJJ22], which was co-led by Benjamin Jarman, Michael Johnson, Jiajie (Jerry) Luo, and myself and co-authored with Mason Porter.[1]

6.1 Introduction

The geographical distribution of resources such as polling sites (i.e., locations where people vote), hospitals, COVID-19 vaccination sites, Department of Motor Vehicles (DMV) locations, and Planned Parenthood clinics is a major factor in the equitability of access to those resources. Consequently, given the locations of a set of resource sites, it is important to quantify their geographical coverage and to identify underserved geographical regions (i.e., "holes in coverage").

A naive approach to quantifying resource coverage is to consider the geographical distances from resource sites by simply calculating the percentage of people who reside within some cutoff distance D of the nearest resource site. This naive approach is common in policy. For example, in March 2021, President Joseph Biden announced a goal of having 90% of the adult United States population within 5 miles (i.e., $D = 5$ miles) of a COVID-19 vaccination

[1]I contributed the main idea for the project (analyzing polling sites with persistent homology with the goal of accounting for real-world factors such as travel time), co-designed the methods with B. Jarman, M. Johnson, and J. Luo, performed the statistical calculations in Section 7.5.6, and wrote the paper with all co-authors.

site [The21]. Additionally, it is required by Indian law that 100% of voters live within 2 km of a polling site [SSH19] (i.e., $D = 2$ km). However, such an approach poses at least two issues:

(1) it requires choosing an arbitrary cutoff distance D; and

(2) using only geographical distance fails to account for many other factors, such as population density and the availability (and facility) of public transportation, that affect ease of access to a resource.

These issues severely limit the utility of this naive approach.

In the present chapter, we use TDA to study holes in resource coverage. A set X of resource sites, with specified latitudes and longitudes, is a point cloud in $\mathbb{R}^2$. We define a metric d that accounts for travel time (see Section 6.2), and we compute the PH of X in the metric space $(\mathbb{R}^2, d)$. One can interpret the PH classes as "holes in coverage." Our TDA approach gives a way to measure and evaluate how equitably a resource is distributed geographically.

Our approach using PH addresses both of the issues (see points (1) and (2) above) of the naive approach that we discussed above. First, PH eliminates the need to choose an arbitrary cutoff distance because one can study holes in coverage at all scales. Second, instead of using geographical distance as our metric, we construct a distance function d that is based on travel times. We also incorporate the waiting time at each resource site by constructing a weighted VR filtration (see Definition 2.2.7) in which we weight vertices using estimates of waiting times at the associated sites. In a city with a high population density or a poor transportation system, the time that is spent waiting at or traveling to a resource site can be a much higher barrier to access than geographical distance [GS03, HK05]. We estimate waiting times by using Global Positioning System (GPS) ping data from mobile phones at the resource sites, and we estimate travel times by using street-network data, per capita car-ownership data, and the Google Maps application programming interface (API) [Goo]. Using these estimates,

we construct a weighted VR filtration in which we weight vertices by our estimates of waiting times; we define the distance between two vertices to be the estimated round-trip travel time between them. Because the weighted VR filtration is stable, small errors in our estimates cause only small errors in the resultant PH [ACG19].

In the present chapter, we examine polling sites as a case study of using PH to study the coverage of resource sites. We restrict our attention to six cities: Atlanta, Chicago, Jacksonville (Florida), Los Angeles[2], New York City (NYC), and Salt Lake City. We use these cities in part because data about them (e.g., car-ownership data) is widely available. Additionally, these cities differ vastly in their demographics and infrastructures, and we can thus compare a variety of different types of cities. Moreover, Atlanta and NYC are both infamous for long waiting times at polling sites, especially in non-White neighborhoods [Fow20, Kan19]. In 2020, some counties in the Atlanta metropolitan area had a mean of 3,600 voters per polling site; the number of polling sites had been cut statewide by 10% since 2013 [Fow20]. In NYC, a mean of 4,173 voters were assigned to each polling site in 2018. As a comparison, in 2004, Los Angeles County and Chicago only had an estimated 1,300 and 725 voters per polling site, respectively [Kan19]. However, Los Angeles is infamous for its traffic [Sch21], and that can hinder voter access to polling sites. Los Angeles and Chicago also differ in the quality of their public transportation, which affects voters' travel times to polling sites. In our investigation, we seek both to compare the coverage of polling sites in our six focal cities and to identify underserved areas within each city.

6.1.1 Related work

One can use tools from geography to study resource accessibility. Pearce et al. [PWB06] used a geographical-information-systems (GIS) approach to study the accessibility of community resources and how it affects health. Hawthorne and Kwan [HK12] used a GIS approach

[2]For Los Angeles, we actually study Los Angeles County. We discuss the reasons for this choice in Subsection 6.2.5.

and a notion of perceived distance to measure healthcare inequality in low-income urban communities. Brabyn and Barnett [BB04] illustrated that there are regional variations in geographical accessibility to general practitioners in New Zealand and that these regional variations depend on how one measures accessibility.

Another motivation for our study of resource-site coverage is the related problem of sensor coverage. Given a set S of sensors in a domain $\Omega \subseteq \mathbb{R}^2$, one seeks to determine if every point in Ω is within sensing range of at least one sensor in S. Typically, each sensor has a fixed, uniform sensing radius r_s. In this case, the problem is equivalent to determining whether or not the domain Ω is covered by balls of radius r_s around each $s \in S$. In [SG07, SG06], de Silva and Ghrist gave homological criteria for sensor coverage. Approaches to study sensor coverage that use computational geometry (specifically, ones that involve the Voronoi diagram for S and the Delauney triangulation of S) were discussed in [LWF03, MKP01].

Our problem is also a coverage problem, but there are important differences. The key conceptual difference is that we consider neighborhoods whose sizes vary with the filtration parameter, rather than neighborhoods of a fixed, uniform radius r_s. Additionally, we do not seek to determine whether or not the balls of any particular radius cover the domain; instead, our goal is to quantify the coverage at all choices of radius and to determine how the holes in coverage evolve as we increase the filtration parameter. Another difference between the present chapter and sensor-coverage problems is that our point cloud represents a set of resource sites (in particular, polling sites), rather than a set of sensors. In a sensor network, pairwise communication between sensors can play a role in whether or not the sensors are fully "connected" to each other (in a graph-theoretic sense) and in determining whether or not a domain is covered [ZH05]. By contrast, communication between resource sites does not play a role in studying access to those resource sites.

Several other papers have applied PH to geospatial data. Feng and Porter [FP21] developed two methods to construct filtrations—one that uses adjacency structures and one that uses the level-set method [OF03] of front propagation—and applied their approaches

to examine geospatial distributions of voting results in the 2016 United States presidential election. They identified "political islands" (i.e., precincts that voted more heavily for a candidate than the surrounding precincts). In [FP20], Feng and Porter used their approaches to study spatial networks. Stolz et al. [SHP16] used PH to examine the geospatial distribution of voting results in the "Brexit" referendum. Hickok et al. [HNP22] used PH to study geospatial anomalies in COVID-19 case-rate data (see also [FHP22]) and vaccination-rate data (see Chapter 5).

6.1.2 Organization

The chapter proceeds as follows. We describe our method in Section 6.2, present the results and analysis of persistence diagrams in Subsection 7.5.6, and conclude and discuss implications, limitations, and potential future directions of our work in Section 6.4. Our code is available at `https://bitbucket.org/jerryluo8/coveragetda/src/main/`.

6.2 Our Construction of Weighted VR Complexes

For each city, we construct a weighted VR filtration (see Definition 2.2.7) in which the point cloud $X = \{x_i\}$ is the set of polling sites in $\mathbb{R}^2$ and the weight w_i of a point x_i is an estimate of the waiting time at the corresponding polling site. We define the radius function

$$r_{x_i}(t) := \begin{cases} -\infty\,, & t < w_{x_i} \\ t - w_{x_i}\,, & \text{otherwise}\,. \end{cases} \tag{6.1}$$

Instead of computing a weighted VR filtration with respect to Euclidean distance, we define a distance function that estimates the mean amount of time that it takes to travel to and from a polling site. With respect to this distance function, the union $\bigcup_i B(x_i, r_{x_i}(t))$ is the set of points y such that the estimated time for an individual at y to vote (including waiting time and travel time in both directions) is at most t.

We construct our distance function as follows. Let x and y be two polling sites. We estimate the expected time for an individual to travel from x to y and back to be

$$\begin{aligned}\tilde{d}(x,y) := C(Z(x)) \min\{t_{\text{car}}(x,y), t_{\text{walk}}(x,y), t_{\text{pub}}(x,y\} \\ + (1 - C(Z(x))) \min\{t_{\text{walk}}(x,y)\,, t_{\text{pub}}(x,y)\}\,,\end{aligned}$$

where $Z(x)$ is the zip code that includes x (a polling site), $C(Z(x))$ is our estimate of the fraction of voting-age people in $Z(x)$ who can travel by car to a polling site, and $t_{\text{car}}(x,y)$, $t_{\text{pub}}(x,y)$, and $t_{\text{walk}}(x,y)$ are our estimates of the expected travel time from x to y and back by car, public transit, and walking, respectively. We estimate $C(Z(x))$ by dividing an estimate of the number of personal vehicles in $Z(x)$ by an estimate of the voting-age population in $Z(x)$; see Subsection 6.2.3. We discuss how we calculate t_{car}, t_{pub}, and t_{walk} in Subsection 6.2.1.

Our definition of $\tilde{d}(x,y)$ captures the cost (in time) to vote. In particular, $\tilde{d}(x,y)$ is an estimate of the mean travel time for somebody who resides in the zip code $Z(x)$ to travel from x to y and back. We assume that all individuals choose the fastest mode of transportation that is available to them. Specifically, we assume that individuals who can travel by car choose the fastest option between driving, taking public transit, and walking. That is, their travel time is $\min\{t_{\text{car}}(x,y), t_{\text{walk}}(x,y), t_{\text{pub}}(x,y\}$. Likewise, we assume that individuals who do not have access to travel by car choose the fastest option between taking public transit and walking. That is, their travel time is $\min\{t_{\text{walk}}(x,y), t_{\text{pub}}(x,y\}$. We estimate that the fraction of a population that has a car is $C(Z(x))$, so the fraction without a car is $1-C(Z(x))$. Therefore, $\tilde{d}(x,y)$ is the (estimated) mean time for a person who resides in zip code $Z(x)$ to travel from x to y and back.

The function $\tilde{d}(x,y)$ is not symmetric (i.e., $\tilde{d}(x,y) \neq \tilde{d}(y,x)$) because $C(Z(x)) \neq C(Z(y))$. However, we need a symmetric function to construct a weighted VR filtration. To construct a symmetric distance function that is based on $\tilde{d}(x,y)$, we define the distance between x and y to be a weighted average of $\tilde{d}(x,y)$ and $\tilde{d}(y,x)$, where we determine the weights from the

populations of the zip codes that include x and y. More precisely, we define the distance between x and y to be

$$d(x,y) := \frac{1}{P}[P_{Z(x)}\tilde{d}(x,y) + P_{Z(y)}\tilde{d}(y,x)]\,, \tag{6.2}$$

where $P_{Z(x)}$, and $P_{Z(y)}$ are the populations of the zip codes $Z(x)$ and $Z(y)$, respectively, and $P := P_{Z(x)} + P_{Z(y)}$ is the sum of the populations of $Z(x)$ and $Z(y)$. With respect to this distance function, the ball $B(x,r)$ is the set of points y such that the expected time for an individual to travel back and forth between x and y is at most r, where the individual is randomly chosen to start at either x or y with probabilities that are weighted by the populations of their associated zip codes.

6.2.1 Estimating travel times

The definition of our distance function (see Equation (6.2)) requires us to estimate the pairwise travel times by car, public transit, and walking between each pair of polling sites. We measure these times in minutes.

We estimate the time that it takes to walk between each pair of polling sites by using street networks, which are available through the OpenStreetMap tool [Ope21], for each of our cities. Using OpenStreetMap, we calculate a shortest path (by geographical distance) between each pair of polling sites. In Figure 6.2.1, we show an example of a shortest path between two polling sites in Atlanta.

Let $L(x,y)$ denote the length (which we measure in meters) of a shortest path (by geographical distance) between polling sites x and y. Our estimate of the walking time (in minutes) from x to y and back is $t_{\text{walk}}(x,y) := 2L(x,y)/v_{\text{walk}}$, where $v_{\text{walk}} = 85.2$ meters per minute is an estimate of the mean walking speed of adult humans [BBH06].

To estimate travel times by car and by public transportation, we use the Google Maps Distance Matrix API [Goo]. Because of budgetary constraints (and the cost of five dollars per thousand API queries), we only use this API to estimate the travel times between each

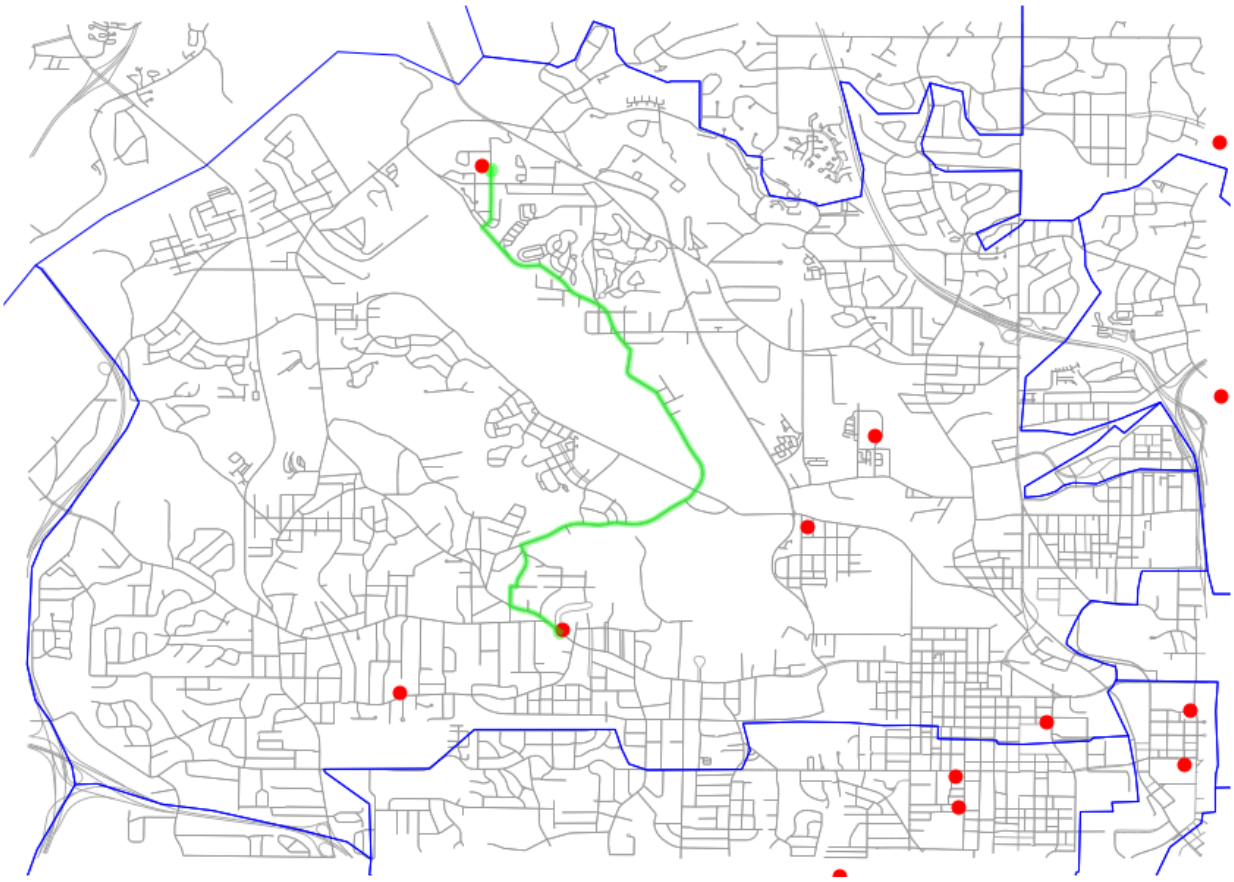


Figure 6.2.1: A shortest path (by geographical distance) between two polling sites in zip code 30314 in Atlanta.

polling site and the 25 geographically closest polling sites to it. We refer to these sites as a polling site's 25 nearest neighbors.

For each of the 25 nearest neighbors, we separately calculate both the time *from* a polling site *to* each neighbor and the time *to* a polling site *from* each neighbor. These two travel times are often different because of differing traffic conditions or other factors. We estimate the remaining pairwise travel times as follows. Let G be the unweighted, undirected graph whose vertices are the polling sites and whose edges connect each vertex to its 25 nearest geographical neighbors.[3] Let G_{car} and G_{pub} be the weighted, directed graphs whose vertices

[3]The relation of being one of a vertex's 25 nearest neighbors is not symmetric. Therefore, the degrees of

and edges are those of G and whose weights are given by the travel times (by car and public transportation, respectively) that we compute using the Google Maps API. The weight of the directed edge from vertex x to vertex y is the travel time from x to y. For two polling sites x and y, we define $t_{\text{car}}(x, y)$ and $t_{\text{pub}}(x, y)$ to be the sums of the lengths of the shortest weighted paths from x to y and y to x in G_{car} and G_{pub}, respectively.

6.2.2 Estimating waiting times

Our weighted VR filtrations have weights at each vertex (i.e., polling site) that are given by an estimate of the mean time spent (i.e., mean waiting time) by a voter at that polling site. In a nationwide study of waiting times at polling sites [CHP19], Chen et al. used smartphone data of hundreds of thousands of voters to estimate waiting times. They also examined potential relationships between waiting times and racial demographics.

We construct our estimates using the congressional district-level estimates in [CHP19] (see their Table C.2) as follows. For each polling site x, we take the mean of the waiting-time estimates for each congressional district that overlaps with the zip code $Z(x)$ that contains x. We believe that this averaging procedure gives a better estimate of the waiting times of individuals who live in the zip code that includes x than simply using the estimate in [CHP19] for the congressional district that includes x.

6.2.3 Estimates of demographic information

We obtain estimates of demographic data at the zip-code level from 2019 five-year American Community Survey data [US]. In particular, we use voting-age population data from Table ACSDT5Y2019.B29001 and vehicle-access data from Table ACSDT5Y2019.B25046.

some vertices are larger than 25.

6.2.4 Polling-site zip codes

Much of our data is at the zip code level, and we treat a polling site's zip code as representative of its local area. Certain polling sites (predominantly government buildings) have their own zip codes, despite their populations of 0. Therefore, we adjust the zip codes of such polling sites to match the zip codes of the directly surrounding areas.

6.2.5 Treatments of Los Angeles and New York City

Because of the oddly-shaped city limits of Los Angeles, which include several holes, we use the entirety of Los Angeles County (except for its islands) when we examine Los Angeles.

Because of the disconnected nature of New York City, we subdivide New York City into three regions (Queens and Brooklyn, Manhattan and the Bronx, and Staten Island) and treat each region separately. We then combine our results for the three regions into a single presentation. For example, we combine the PDs into a single PD for all of New York City.

6.3 Results

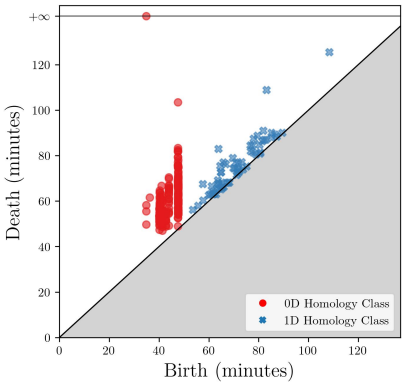
We compute the PH of the weighted VR filtrations of Section 6.2 for Atlanta, Chicago, Jacksonville, Los Angeles County, NYC, and Salt Lake City. We show their PDs in Figure 6.3.1. A homology class that dies at filtration-parameter value t represents a "hole in coverage" that persists until time t. One can interpret this to mean that somebody who lives in a "hole in coverage" that dies at t needs more than t minutes (including both waiting time at a polling site and travel time back and forth to the site) to cast a vote. We consider both 0D and 1D homology classes in our analysis. 1D homology classes represent "holes in coverage" bounded by closed paths, while 0D homology classes represent holes between different connected regions of coverage. In our analysis, we do not use the homology-class birth values because we view them as irrelevant for this application. The reason for this

City	Homology Dimension	Median (minutes)	Variance (minutes)
Atlanta	0	59.8	75.9
	1	74.0	140.2
Chicago	0	53.1	30.2
	1	65.0	57.2
Jacksonville (Florida)	0	42.8	75.7
	1	57.5	331.9
Los Angeles County	0	59.5	54.1
	1	74.3	74.3
New York City	0	55.0	55.3
	1	68.0	104.9
Salt Lake City	0	82.6	37.3
	1	99.1	23.0

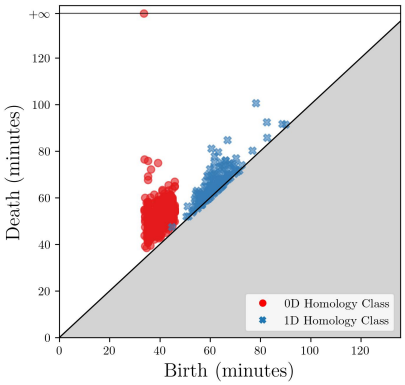
Table 6.3.1: The medians and variances of the homology-class death values for each city. (As we discussed in the main text, we consider Los Angeles County, rather than only the city of Los Angeles.)

is that a birth value indicates only the filtration-parameter value at which a coverage hole materializes. In Figure 6.3.2, we show a box plot of the distribution of homology-class death values for each city. Larger death values suggest that a city may have worse coverage, and a wider distribution of death values suggests that there may be more variation in polling-site accessibility within a city. In Table 6.3.1, we show the medians and variances of the 0D and 1D homology-class death values for each city.

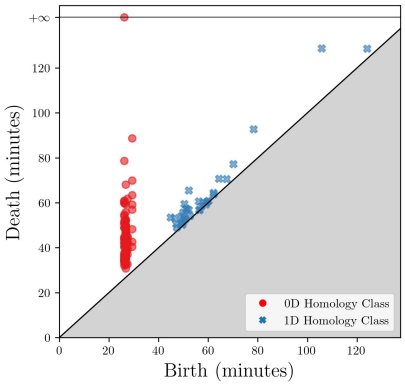
We compare the coverages of the cities by examining the death values in the PDs. For example, in the PDs for Atlanta and Chicago in Figure 6.3.1, we see that Atlanta's homology



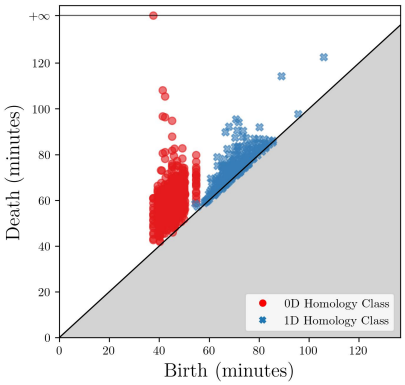
(a) Atlanta



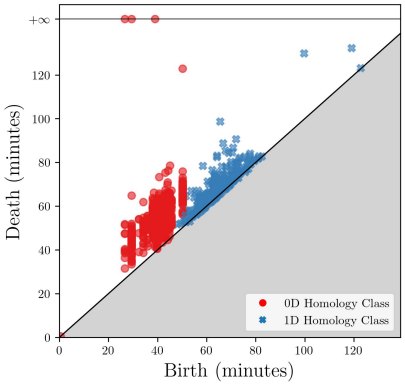
(b) Chicago



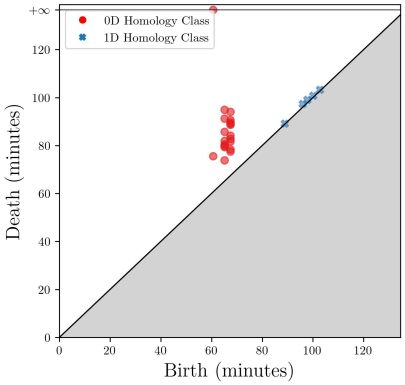
(c) Jacksonville (Florida)



(d) Los Angeles County



(e) New York City



(f) Salt Lake City

Figure 6.3.1: Our PDs for each city for the PH of the weighted VR complexes that we defined in Section 6.2.

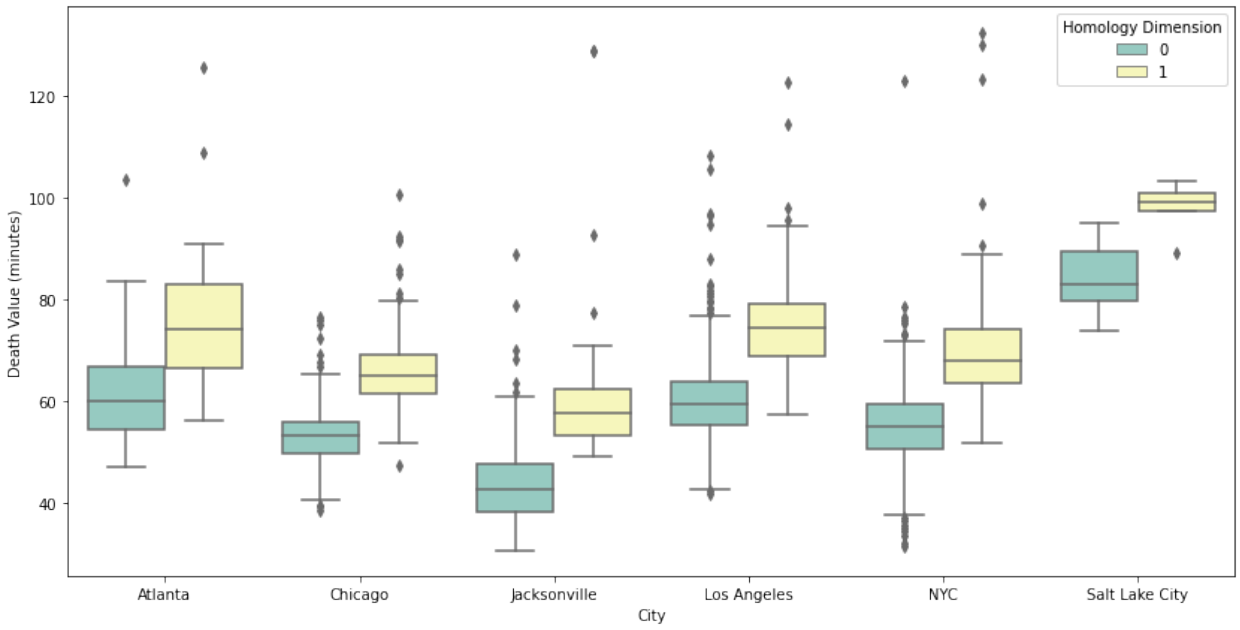
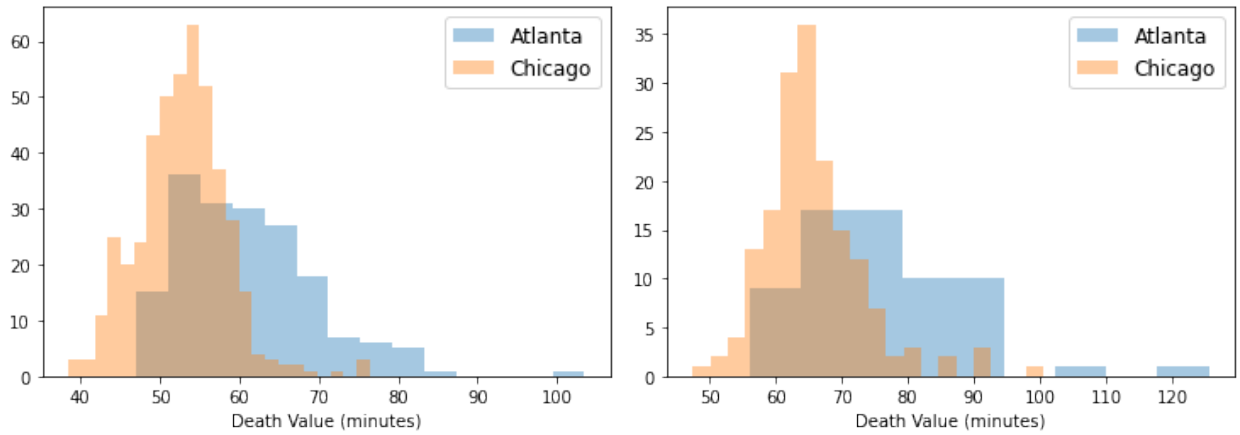


Figure 6.3.2: Box plots of the death values of the 0D and 1D homology classes for each city.

classes tend to die later in the filtration than Chicago's homology classes. We can also see this in Figure 6.3.2, in which we show box plots of the death values for each city, and in Figure 6.3.3, in which we plot the distributions of death values for Atlanta and Chicago. Our PDs and visualizations of summary statistics suggest that Chicago has better polling-site coverage than Atlanta.

We use the death simplices to locate and visualize holes in polling-site coverage. We interpret the death simplex of a homology class as the "epicenter" of the associated coverage hole because the death simplex represents the last part of the hole to be covered. For example, the death simplex of a 0D homology class is an edge between two polling sites; we interpret the homology class as a hole in coverage between those two polling sites. The death simplex of a 1D homology class is a triangle that is the convex hull of three polling sites; we interpret the homology class as a hole in coverage between those three polling sites. In Figures 6.3.4 and 6.3.5, we show the death simplices with the largest death values.[4] For

[4]More precisely, for each city and each homology dimension (0 and 1), we show the death simplices whose



(a) 0D homology classes

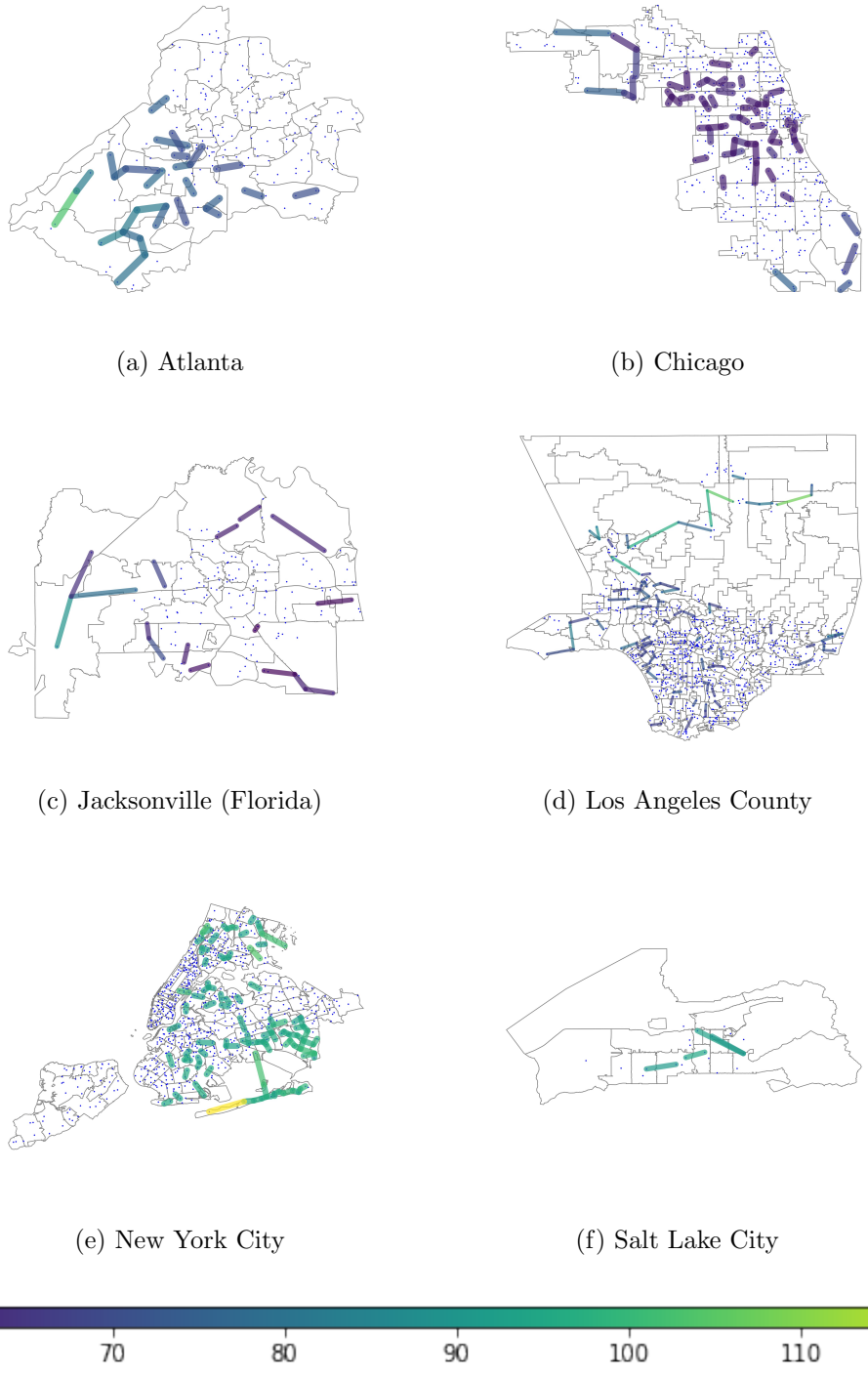
(b) 1D homology classes

Figure 6.3.3: Histograms of the death values of the 0D and 1D homology classes for Atlanta and Chicago.

example, consider panels (a) and (b) of Figure 6.3.4 and Figure 6.3.5, in which we show the 0D and 1D homology-class death simplices for Atlanta and Chicago. The areas of lowest coverage (i.e., the areas that contain the death simplices with the largest death values) in Atlanta tend to be in the southwest, whereas the areas of lowest coverage in Chicago tend to be in the northwest and southeast. There is one homology class in Atlanta that has a significantly larger death filtration value than the other classes in Atlanta and any of the classes in Chicago. This homology class represents a 1D hole in coverage in southwest Atlanta (see Figure 6.3.5a).

---

death values have a z-score of at least 1. We calculate the z-score as follows. Let d be the death value of a q-dimensional homology class (where $q = 0$ or $q = 1$) for city C. The z-score of d is $z = (d - \mu_{C,q})/\sigma_{C,q}$, where $\mu_{C,q}$ and $\sigma_{C,q}$ are the mean and standard deviation of the distribution of death values of the q-dimensional homology classes for city C.



(a) Atlanta

(b) Chicago

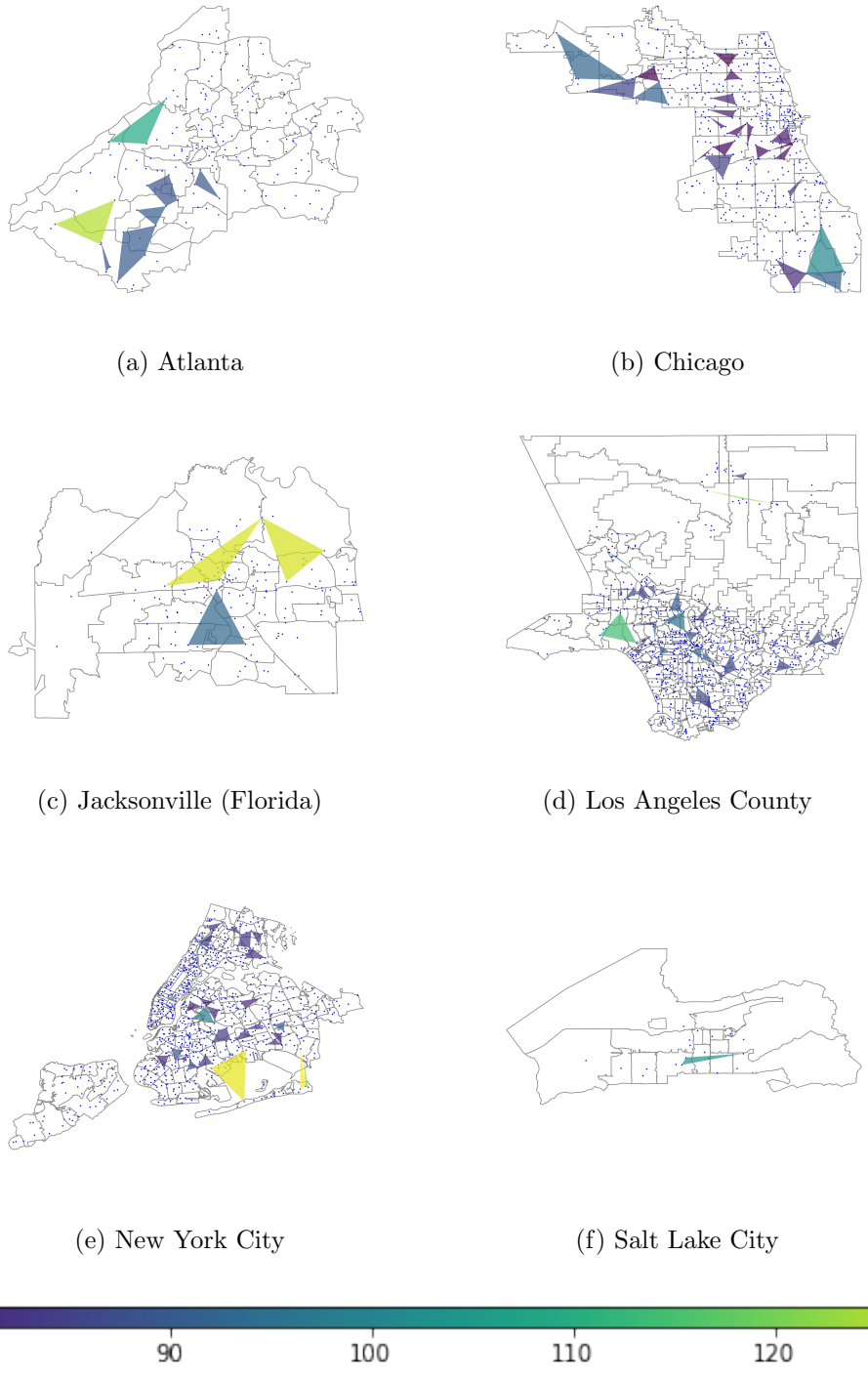
(c) Jacksonville (Florida)

(d) Los Angeles County

(e) New York City

(f) Salt Lake City

Figure 6.3.4: Death simplices for the 0D homology classes with the largest death values. The colors correspond to the death values.



(a) Atlanta

(b) Chicago

(c) Jacksonville (Florida)

(d) Los Angeles County

(e) New York City

(f) Salt Lake City

Figure 6.3.5: Death simplices for the 1D homology classes with the largest death values. The colors correspond to the death values.

6.4 Conclusions and Discussion

6.4.1 Summary

We showed that PH is a helpful approach to study accessibility and equitability. It allows one to examine holes in resource coverage with respect to an appropriate choice of "distance," which one constructs to incorporate important features of a problem of interest. The distance can be based on geography, time, or something else. In the present chapter, we used PH to study and quantify holes in polling-site coverage in six United States cities (technically, in five cities and in Los Angeles County). For each city, we constructed a filtration in which a homology class that dies at time t represents a geographical region in which it takes more than t minutes to cast a vote (including both travel time and waiting time). We interpreted the death simplex of a homology class as the location of the corresponding hole in resource coverage. The information in the PH allowed us both to compare the accessibility of voting across our chosen cities and to determine the locations of the coverage holes within each city.

A key benefit of our use of PH is that it enabled us to identify holes in coverage at all time scales. Moreover, it allowed us to use a distance that we designed for the problem at hand, rather than merely using geographical distance, which does not capture important factors in resource accessibility [BB06]. We based our distance function on estimates of travel time, which is more reasonable and accurate than geographical distance for capturing resource accessibility [PWB06].

6.4.2 Limitations

To conduct our study, we needed to estimate a variety of quantities (see Section 6.2), including travel times, waiting times, and demographic information. We also made several simplifications because of computational and monetary constraints. We now discuss some issues that are important to address before attempting to incorporate our approach into

policy-making.

One limitation of our study is our estimation of travel times. As we discussed in Subsection 6.2.1, we computed travel times using the Google Maps API. Because of monetary constraints, we only computed a subset of the travel times and used a graph-based estimate to determine the others. Additionally, we computed each travel time between polling sites only once. Computing more precise estimates of travel times is important for better capturing the accessibility of polling sites. One way to do this is to compute the same travel time multiple times across different days and times of day and take an average.

Another limitation of our study is the granularity of our data. As we discussed in Subsection 6.2.2, our waiting-time data is at the scale of congressional districts. Because there is heterogeneity in the waiting times at different polling sites in the same congressional district, it is important to obtain finer-grained data for the waiting times at polling sites. Having finer-grained waiting times (e.g., if possible, procuring an estimated waiting time for each polling site) would improve our ability to capture voting accessibility.

We also made several topological approximations. For each city, we worked with a weighted VR filtration, which approximates a weighted Čech filtration, which in turn approximates the nested set $\{\bigcup B(x_i, r_{x_i}(t))\}_{t\in\mathbb{R}}$ of spaces, where $\{x_i\}$ is a set of polling site locations and $r_{x_i}(t)$ is the radius function that we defined in Section 6.2. The nested set of spaces is directly relevant to our application, as the holes in $\bigcup B(x_i, r_{x_i}(t))$ are the true holes in polling-site coverage. We made our approximations, which are standard in TDA and are well-justified (see Section 2.2) [OPT17], to reduce computational cost. However, the convexity condition of the Nerve Theorem (see Theorem 2.2.3 in Chapter 2), which is what justifies the approximation of $\bigcup B(x_i, r_{x_i}(t))$ by a weighted Čech complex, is not guaranteed to be satisfied for all times t. The Nerve Theorem implies that the weighted Čech complex is homotopy-equivalent to $\bigcup B(x_i, r_{x_i}(t))$ whenever the balls $B(x_i, r_{x_i}(t))$ are convex. This condition always holds in Euclidean space, but it is not guaranteed to hold in the space that we defined in Section 6.2. Homotopy-equivalence is important because homotopy-equivalent

spaces have the same homology and thus have the same "holes."

Finally, our approach only detects holes in the convex hull of a set of resource sites. Although this may be inconsequential if resource sites are sufficiently spread out geographically, it can be problematic if the resource sites are overly concentrated near a few locations. One way to address this issue is to incorporate the city boundaries into the construction of the filtrations. This would help capture holes in coverage in regions that lie outside the convex hull of the resource sites and to help identify the filtration-parameter value t at which an entire city is covered by the balls $B(x_i, r_{x_i}(t))$.

6.4.3 Future work

As we discussed in Subsection 6.4.2, we made several topological approximations of the object of interest, which is the nested set $\{\bigcup B(x_i, r_{x_i}(t))\}_{t\in\mathbb{R}}$ of spaces. Instead of using a weighted VR filtration, one can construct a more direct approximation of $\{\bigcup B(x_i, r_{x_i}(t))\}_{t\in\mathbb{R}}$ as follows. One can first discretize a city by imposing a grid onto it, and one can then construct the filtered cubical complex that is induced by the distance to the nearest polling site. However, this is much more computationally expensive than our approach in the present chapter, and it would also entail many more travel-time queries (which cost money) than in the present chapter.

It is also important to incorporate city boundaries into the construction of filtrations. One way to do this is as follows. Let $x_1, \ldots, x_n$ denote the resource sites, and let $y_1, \ldots, y_m$ denote the points that one obtains by discretizing a city boundary. One can extend our distance function Equation (6.2) by defining[5]

$$d(x_i, y_j) := \frac{2}{P}[P_{Z(x_i)}\tilde{d}(x_i, y_j) + P_{Z(y_j)}\tilde{d}(y_j, x_i)]\,, \tag{6.3}$$

[5]The factor of 2 comes from the fact that x_i is a resource site but y_j is not.

where P, Z, and $\tilde{d}$ are as in Equation (6.2) and

$$d(y_i, y_j) = \begin{cases} 0\,, & y_i \text{ and } y_j \text{ are adjacent points of the discretized city boundary} \\ \infty\,, & \text{otherwise}\,. \end{cases} \tag{6.4}$$

At each filtration-parameter value, the simplicial complex that one constructs using the distance Equation (6.2) with the extensions Equation (6.3) and Equation (6.4) includes both the points that one obtains by discretizing the boundary and the edges that connect adjacent boundary points. The largest death value is then the filtration-parameter value t that corresponds to when an entire city is covered by the balls $\{B(x_i, r_{x_i}(t))\}$ (i.e., when there are no holes in coverage).

We used death simplices to locate the holes in coverage, but there are many other approaches that one could take. By calculating minimal generators [LTH21], one can identify representative cycles that encircle the holes. HyperTDA [BYM22] is a technique to analyze the structure of minimal generators by constructing a hypergraph and applying hypergraph centrality measures and community detection methods. This approach might provide insights into the spatial structure of the minimal generators. Another approach involves using Decorated Merge Trees (DMTs) [CHM22] to locate 1D holes in coverage. The cluster of points (polling sites) that contain a 1D hole in coverage is the set of points (polling sites) corresponding to the leaves in the DMT descended from the birth point of that hole. Although these approaches are beyond the scope of the present chapter, it would be interesting to explore them in future work.

Although we explored a specific case study (namely, the accessibility of polling sites), it is also relevant to conduct similar investigations for other resources, such as public parks, hospitals, vaccine distribution centers, grocery stores, Planned Parenthood clinics, and Department of Motor Vehicle (DMV) locations. One can use similar data to construct a filtration, although it may be necessary to modify the choices of distance and weighting. One can also use ideas from mobility theory [BBG18] to help construct suitable distances and weight-

ings. For example, all DMV offices offer largely the same services, so it seems reasonable to assume that people will go to their nearest office. Therefore, in a study of DMV accessibility, it seems appropriate to use travel time as a distance function, just as we did in our analysis of polling sites. However, in other applications, it is not reasonable to use travel time alone as a distance function. For example, different grocery stores may offer different products at different prices, so travel time alone may not be appropriate as a choice of distance function. Additionally, although waiting time is a significant factor for investigating the coverage of polling sites, there are many applications for which it does not make sense to incorporate waiting time. For example, the time that is spent in a public park or recreation center is typically not a barrier to access. In applications for which waiting time is not an accessibility factor, it seems more appropriate to use a standard VR filtration than a weighted VR filtration. With salient modifications (such as the ones that we described in this subsection and in Subsection 6.4.2), we can apply our approach to many other types of resource sites.

Acknowledgements

We thank Chris Anderson and Renata Turkes for helpful comments and discussions.

CHAPTER 7

An Intrinsic Approach to Scalar Curvature Estimation for Point Clouds

This chapter is adapted from a paper with Andrew Blumberg that is in preparation.

7.1 Introduction

Curvature, which measures the extent to which a Riemannian manifold deviates from being "flat," is a generalization of the use of the second derivative to measure the extent to which a curve pulls away from the tangent line at a point. There are several different notions of curvature in Riemannian geometry. The focus of this chapter is *scalar curvature*, which quantifies the curvature at a point by a scalar in $\mathbb{R}$. One way to think of scalar curvature is that it is proportional to the "average" *sectional curvature*, which assigns a curvature value to every plane in the tangent space at a point by considering the curvature of the 2D submanifold that is determined by the plane. On a surface, scalar curvature is equal to twice the Gaussian curvature.

The purpose of this chapter is to study the problem of estimating the scalar curvature of a manifold given a finite sample $X \subset M$, which we assume consists of independent draws from some (possibly nonuniform) probability density function $\rho : M \to \mathbb{R}_+$. Our estimator is based on the fact that scalar curvature characterizes the growth rate of the volume of a geodesic ball $B^M(x, r)$ as r increases. As $r \to 0$, the scalar curvature $S(x)$ at $x \in M$ has the

following relationship to geodesic-ball volume:

$$\frac{\mathrm{vol}(B^M(x,r))}{v_n r^n} = 1 - \frac{S(x)}{6(n+2)} r^2 + \mathcal{O}(r^4)\,, \tag{7.1}$$

where n is the dimension of the manifold, v_n is the volume of a unit Euclidean n-ball, and $v_n r^n$ is the volume of a Euclidean n-ball of radius r. We proceed by computing maximum-likelihood estimators for the volumes on the left side of equation (7.1) and fitting a quadratic function to approximate $S(x)$. Our estimate $\hat{S}(x)$ is $-6(n+2)\hat{C}(x)$, where $\hat{C}(x)$ is the quadratic coefficient of the fitted curve.

Our first main theorem shows that our scalar-curvature estimator is asymptotically stable. That is, small errors in distance measurement (e.g., from geodesic-distance estimation) cause only small errors in our scalar-curvature estimates. This is especially important so that one can accurately estimate scalar curvature in real-world data sets, which are invariably noisy.

Theorem A (Theorem 7.3.1)**.** *Let M be a compact Riemannian manifold that is equipped with a probability measure ρ that has full support. Let $\{X_k\}_{k\in\mathbb{N}}$ be a sequence of finite samples, with $|X_k| \to \infty$, that are drawn from M according to ρ and equipped with metrics d_k such that*

$$\max_{(x,y)\in X_k \times X_k} |d_k(x,y) - d(x,y)| \to 0 \qquad \text{as } k \to \infty\,,$$

where $d(x,y)$ is the geodesic distance between x and y. Then for a suitable sequence of radius sequences and any sequence $\{x_k\}$ of points with $x_k \in X_k$, we have $|\hat{S}[d_k](x_k) - \hat{S}[d](x_k)| \to 0$ in probability as $k \to \infty$.

Remark 7.1.1. In fact, we can extract an effective stability bound from the proof of Theorem 7.3.1; the distance between curvature estimators is controlled by an explicit bound that involves an additive term based on the radius sequence and a term that is controlled by the discrepancy $\delta_k := \max_{(x,y)\in X_k\times X_k} |d_k(x,y) - d(x,y)|$.

We can then establish our second main theorem, which establishes that our scalar curvature estimate $\hat{S}(x)$ converges to the true scalar curvature $S(x)$ under certain asymptotic conditions as $|X_k| \to \infty$.

Theorem B (Theorem 7.4.3)**.** *Under the same hypotheses as the preceding theorem, for a suitable sequence of radius sequences,* $|\hat{S}[d_k](x_k) - S(x_k)| \to 0$ *in probability as* $k \to \infty$*, where* $\{x_k\}$ *is any sequence of points such that* $x_k \in X_k$ *for each* k*.*

We test our scalar-curvature estimator on point clouds that are sampled from several different manifolds (see Section 7.5). Broadly, the experiments demonstrate that our method is accurate on manifolds of constant scalar curvature, especially on low-dimensional manifolds. Additionally, our method is robust with respect to additive isotropic noise.

The primary limitation of our method is that we typically cannot accurately estimate scalar curvature on regions of the manifold where the scalar curvature has high variation or where it achieves a minimum or maximum. For example, in our experiments on the torus and hyperboloid, we find that we cannot accurately estimate curvature near points where scalar curvature is minimized. However, our results are qualitatively correct even when the pointwise estimates are inaccurate. On the torus and the hyperboloid, we correctly detect the scalar-curvature sign and relative changes in scalar curvatures across the surfaces.

Applications

We estimate scalar curvature using only metric data (i.e., pairwise geodesic distances between points that are sampled from on a manifold). Consequently, our estimator can be applied to any finite metric space. We expect that our estimator can be used to inform a choice of low-dimensional embedding space. For example, if one finds that the scalar curvature of a data set is negative (respectively, positive) everywhere, then this would suggest embedding the points into hyperbolic space (respectively, a sphere). In recent years, there has been much research on non-Euclidean embeddings, such as hyperbolic embeddings [KLB20, SDG18, NK17].

Another application is the generalization of curvature to metric spaces that do not obviously come from manifolds. A notable example is given by a network with the shortest-path metric. This yields a definition of discrete scalar curvature that is defined on the vertices of

a network.

Related Work

There are several other papers on curvature estimation. Sritharan et al. [SWH21] developed a different method to estimate scalar curvature by using the second fundamental form and the Gauss–Codazzi equation. However, their method requires an embedding of the points in Euclidean space. Furthermore, it is sensitive to noise because it involves tangent-space estimation. In one experiment, in which points were sampled from a Klein bottle, their method did not recover the correct sign for the scalar curvature after only a small amount of Gaussian noise was added (standard deviation $\sigma = .01$). By contrast, we are able to obtain higher accuracy on noisy data sets.

Bhaskar et al. [BMF22] defined "diffusion curvature," which is a new (unsigned) measure of local curvature for point clouds. Although diffusion curvature is not the same as Gaussian or scalar curvature, numerical experiments in [BMF22] suggest that it is correlated with Gaussian curvature. However, unlike Gaussian and scalar curvature, diffusion curvature is always positive, so it cannot be used to infer whetherscalar curvature is positive or negative. By contrast, our scalar-curvature estimates are signed, so our method can be used to distinguish between regions of positive and negative curvature.

Chazal et al. [CCL09] showed that curvature measures (which are distinct from curvature) can be estimated stably. However, as in [SWH21], their method requires an embedding of the points in Euclidean space. Moreover, their method is not feasible for point clouds in high dimensions because it requires computing and storing the boundaries and intersections of a set of balls in the ambient space; Chazal et al. implemented and tested their method only in $\mathbb{R}^3$. By contrast, the accuracy and computational complexity of our method in the present chapter depends only on the intrinsic dimension of the manifold; it does not depend on the dimension of the ambient space.

Topological data analysis can also be used to detect curvature. Bubenik et al. [BHP20] used persistent homology to classify point clouds by the curvature of the constant-curvature surface from which they were sampled. However, the scope of [BHP20] was limited to Gaussian curvature of constant-curvature surfaces.

Lastly, we note that there is an important relationship to discrete network curvature [SSG18]. Our scalar-curvature estimator can be applied to networks with the shortest-path metric. There are two other definitions of discrete scalar curvature for networks, both of which are defined as "contractions" of discrete Ricci curvature [SGR15, SJS17].[1] These are justified by the fact that scalar curvature is the trace of Ricci curvature. However, it has not been proven that either notion of discrete scalar curvature converges to the scalar curvature of the manifold when the network is a geometric network on a manifold.

Organization

We discuss our method for estimating scalar curvature in Section 7.2. We prove stability (Theorem 7.3.1) in Section 7.3 and convergence (Theorem 7.4.3) in Section 7.4. Finally, we discuss our numerical experiments in Section 7.5.

7.2 Estimating scalar curvature via geodesic ball-volume estimation

Suppose that we are given a *distance matrix* d_X, which is a matrix whose (i, j)th entry is the distance between the ith and jth points of a point cloud X that consists of N points. By a slight abuse of notation, we will write $d_X(x, y)$ to denote the distance between points $x \in X$ and $y \in X$. We assume that

[1] More precisely, the discrete scalar curvature at a node is defined to be the sum of its adjacent edges' discrete Ricci curvature. Sandhu et. al [SGR15] defined scalar curvature at a vertex as the contraction of Ollivier–Ricci curvature. Sreejith et. al [SJS17] defined scalar curvature as the contraction of Forman–Ricci curvature.

1. (X, d_X) is a metric subspace of an unknown Riemannian manifold (M, g) of unknown dimension n and

2. X is sampled randomly from an unknown probability density function $\rho : M \to \mathbb{R}_+$.

Let d denote the geodesic distance on M. Importantly, we do not assume that we are given coordinates for the point cloud X; we assume only that we have the distance matrix d_X. However, it is possible to begin with a point cloud X (instead of its distance matrix d_X), from which one can estimate geodesic distances using, for example, the graph-approximation technique of Tenenbaum et al. [TdL00, BSL00].

We summarize our scalar-curvature estimation method in Figure 7.2.1. To estimate the scalar curvature $S(x)$ at a point $x \in X$, the idea of our approach is to estimate $\mathrm{vol}(B^M(x, r))$ for a sequence of increasing r and then estimate $S(x)$ by fitting a quadratic polynomial to the estimated ball-volume ratios $\mathrm{vol}(B^M(x, r))/(v_n r^n)$.

7.2.1 Maximum-likelihood estimator of ball volume

For a given radius r and a point $x \in X$, we estimate $\mathrm{vol}(B^M(x, r))$ as follows. Let N be the number of points in X, and let $N[d_X](x, r)$ denote the number of points in $B^M(x, r) \cap (X \setminus \{x\})$. That is,

$$N[d_X](x, r) := \left|\{y \in X \setminus \{x\} \mid d_X(x, y) \leq r\}\right| .$$

When the metric d_X is clear from context, we omit it from the notation and write $N(x, r)$. Let $\mu_\rho(x, r)$ denote the mean density within $B^M(x, r)$. That is,

$$\mu_\rho(x, r) := \frac{1}{\mathrm{vol}(B^M(x, r))} \int_{z \in B^M(x,r)} \rho(z) dV ,$$

where dV is the volume form on M. When $\rho(x) \equiv \rho$ is constant, $\mu_\rho(x, r) = \rho$. In Section 7.2.4, we discuss a method to estimate $\mu_\rho(x, r)$ empirically, without prior knowledge of $\mathrm{vol}(B^M(x, r))$.

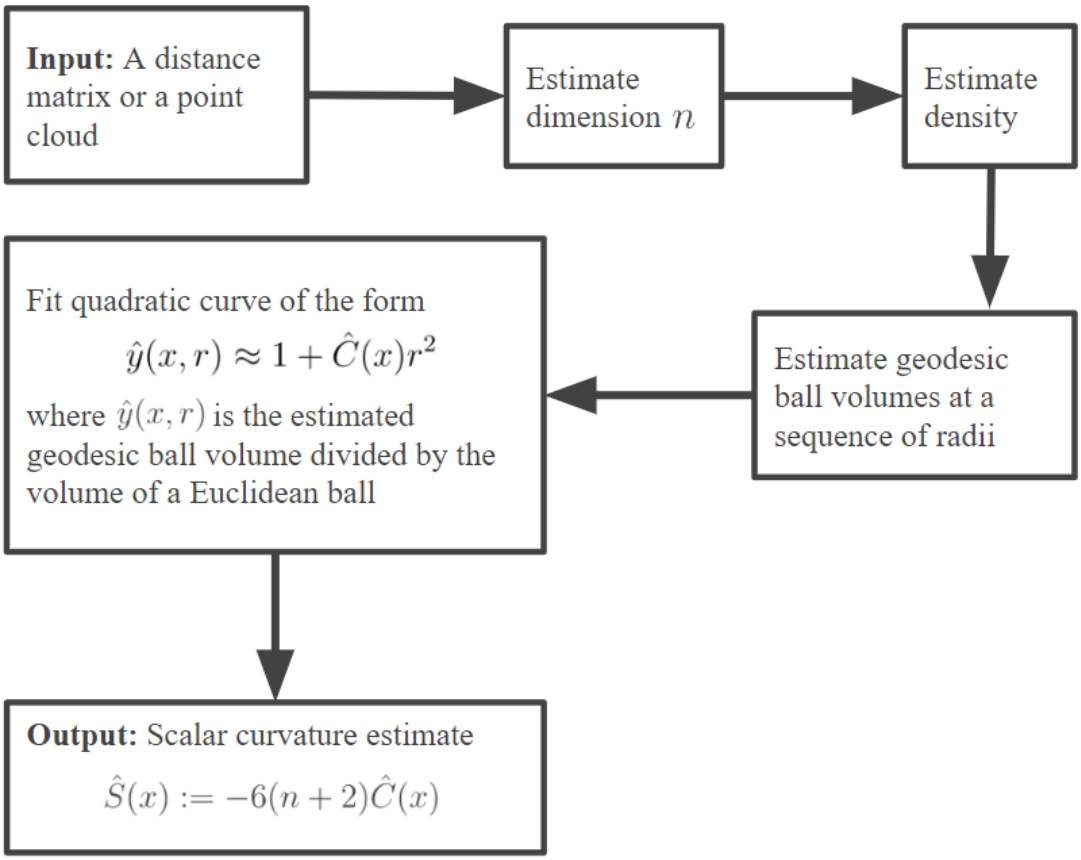


Figure 7.2.1: The pipeline for our scalar-curvature estimation method.

Our likelihood function for $\mathrm{vol}(B^M(x,r))$ is

$$\begin{aligned} L(v) &= \mathbb{P}[N(x,r) \mid \mathrm{vol}(B^M(x,r)) = v] \\ &= \binom{N-1}{N(x,r)} \Big(\mu_\rho(x,r)v\Big)^{N(x,r)} \Big(1-\mu_\rho(x,r)v\Big)^{N-1-N(x,r)} \end{aligned}$$

because the random variable $N(x,r)$ is a binomial random variable with $N-1$ trials and success probability $\mu_\rho(x,r) \cdot \mathrm{vol}(B^M(x,r))$. Solving $0 = L'(v)$, we find that the maximum-likelihood estimator is

$$v_* = \frac{N(x,r)}{(N-1)\mu_\rho(x,r)}\,. \tag{7.2}$$

The expectation of v_* is

$$\mathbb{E}[v_*] = \frac{\mathbb{E}[N(x,r)]}{(N-1)\mu_\rho(x,r)} = \mathrm{vol}(B^M(x,r)). \tag{7.3}$$

7.2.2 Dimension estimation

Our scalar-curvature estimation method requires an estimate $\hat{n} \in \mathbb{N}$ of the manifold dimension n; there are myriad methods to do this [CS16]. One method to estimate dimension is the maximum-likelihood method of Levina and Bickel [LB04], which requires only the distance matrix d_X as input. (See Section 7.5.2 for details.) When the distance matrix d_X is not clear from context, we denote our dimension estimate by $\hat{n}[d_X]$. We assume that $\hat{n} = n$ in our theoretical results (Sections 7.3 and 7.4), and we have $\hat{n} = n$ for all data sets in our numerical experiments (Section 7.5).

7.2.3 Density estimation

In Equation 7.2, the mean density $\mu_\rho(x,r)$ in the ball must be estimated empirically. To do so, we first empirically estimate the density at each point $z \in X$. One method for doing so is kernel density estimation (KDE) on a manifold [OG09], which requires only the distance matrix d_X and an estimate $\hat{n}$ of the manifold dimension as input.

Remark 7.2.1. We denote a choice of density estimator by $\hat{\rho}$, and we denote our density

estimate at $z \in X$ by $\hat{\rho}[d_X, \hat{n}](z)$. For example, in our numerical experiments in Section 7.5, the density estimator $\hat{\rho}$ is a kernel density estimator with either a Gaussian or biweight kernel. If d_X and $\hat{n}$ are clear from context, we omit them and write $\hat{\rho}(z)$.

After we compute our pointwise-density estimates $\hat{\rho}(z)$ for all $z \in X$, we calculate an estimate $\widehat{\mu_\rho}[\hat{\rho}](x, r)$ of the mean density $\mu_\rho(x, r)$ within $B^M(x, r)$. We define

$$\widehat{\mu_\rho}[\hat{\rho}](x, r) := \begin{cases} \left(\frac{1}{N(x,r)} \sum_{z \in B^M(x,r) \cap (X \setminus \{x\})} 1/\hat{\rho}(z) \right)^{-1}, & N(x, r) > 0 \\ \hat{\rho}(x)\,, & N(x, r) = 0\,. \end{cases} \tag{7.4}$$

We write $\widehat{\mu_\rho}[\rho](x, r)$ when $\hat{\rho}(x) = \rho(x)$ for all $x \in X$.

Notably, our estimate $\widehat{\mu_\rho}[\hat{\rho}](x, r)$ is not the sample mean of $\{\hat{\rho}(z) \mid z \in B^M(x, r) \cap (X \setminus \{x\})$. The sample mean $\frac{1}{N(x,r)} \sum_z \hat{\rho}(z)$ is an overestimate of $\mu_\rho(x, r)$ because points with high density are overrepresented in the sample $B^M(x, r) \cap (X \setminus \{x\})$.

Remark 7.2.2. When r is small, we have $\mu_\rho(x, r) \approx \rho(x) \approx \hat{\rho}(x)$. Indeed, one can show that $\mu_\rho(x, r) \to \rho(x)$ as $r \to 0$. However, in our numerical experiments in Section 7.5, our scalar-curvature estimation method sometimes requires us to estimate $\mu_\rho(x, r)$ when r is not small. Informally, what we show in Lemma 7.2.3 is that $1/\widehat{\mu_\rho}[\hat{\rho}](x, r)$ is a good approximation to $1/\mu_\rho(x, r)$ even for large r. (Estimating the maximum-likelihood estimator of Equation 7.2 requires us to estimate the reciprocal $1/\mu_\rho(x, r)$.) In our experiments (Section 7.5), we observe significant empirical improvement from using Equation 7.4 to estimate $\mu_\rho(x, r)$ instead of using $\hat{\rho}(x)$ to estimate $\mu_\rho(x, r)$. This observation holds even when the data is uniformly sampled because $\widehat{\mu_\rho}[\hat{\rho}](x, r)$ averages the empirical densities (which may differ from the ground-true density) within the ball.

Lemma 7.2.3. If X is a point cloud sampled from the probability density function $\rho : M \to \mathbb{R}_+$, then

$$\mathbb{E}\left[\frac{1}{\widehat{\mu_\rho}[\rho](x, r)}\right] = \frac{1}{\mu_\rho(x, r)}$$

for all $x \in X$ and $r > 0$.

Proof. The quantity $\frac{1}{\widehat{\mu_\rho}[\rho](x,r)}$ is the sample mean of $1/\rho(z)$ for $z \in B^M(x,r) \cap (X \setminus \{x\})$. Therefore,

$$\mathbb{E}\Big[\frac{1}{\widehat{\mu}_\rho[\rho](x,r)}\Big] = \mathbb{E}\Big[\frac{1}{\rho(z)}\Big]\,,$$

where z is a point that is conditioned to lie in $B^M(x,r)$. The probability density function for z is

$$\psi(z) := \frac{\rho(z)}{\int_{w \in B^M(x,r)} \rho(w) dV}\,. \tag{7.5}$$

Therefore,

$$\mathbb{E}\Big[\frac{1}{\rho(z)}\Big] = \int_{z \in B^M(x,r)} \frac{1}{\rho(z)} \psi(z) dV = \frac{\mathrm{vol}(B^M(x,r))}{\int_{w \in B^M(x,r)} \rho(w) dV} = \frac{1}{\mu_\rho(x,r)}\,. \tag{7.6}$$

□

7.2.4 Empirical approximation of the maximum-likelihood estimator

For a given $x \in X$ and radius $r > 0$, we define our estimate of $\mathrm{vol}(B^M(x,r))$ to be

$$\widehat{\mathrm{vol}}[d_X, \hat{\rho}](x,r) := \frac{N[d_X](x,r)}{(N-1)\widehat{\mu}_\rho[\hat{\rho}](x,r)}\,, \tag{7.7}$$

where $\widehat{\mu}_\rho[\hat{\rho}](x,r)$ is defined as in Eq. 7.4. We write $\widehat{\mathrm{vol}}[d_X, \rho](x,r)$ when $\hat{\rho}(x) = \rho(x)$ for all $x \in X$.

The quantity $\widehat{\mathrm{vol}}[d_X, \hat{\rho}](x,r)$ is an approximation of the true maximum-likelihood estimator v_* in Eq. 7.2. An equivalent formula for $\widehat{\mathrm{vol}}[d_X, \hat{\rho}](x,r)$ is

$$\widehat{\mathrm{vol}}[d_X, \hat{\rho}](x,r) = \frac{\sum_{z \in B^M(x,r) \cap (X \cap \{x\})} \frac{1}{\hat{\rho}(z)}}{(N-1)}\,. \tag{7.8}$$

Lemma 7.2.4. If X is a finite point cloud that is sampled from the probability density function $\rho : M \to \mathbb{R}_+$, then

$$\mathbb{E}\Big[\widehat{\mathrm{vol}}[d_X, \rho](x,r)\Big] = \mathrm{vol}(B^M(x,r))$$

for all $x \in X$ and $r > 0$.

Proof. Let N be the number of points in X. For all $k \in \{0, \ldots, N-1\}$,

$$\mathbb{E}\Big[\widehat{\mathrm{vol}}[d_X, \rho](x,r)\Big| N(x,r) = k\Big] = \frac{k}{N-1}\mathbb{E}[1/\rho(z)] \tag{7.9}$$

by Equation 7.8, where z is randomly drawn according to the probability density function $\psi(z)$ defined in Equation 7.5. Substituting Equation 7.6 into Equation 7.9 yields

$$\mathbb{E}\Big[\widehat{\mathrm{vol}}[d_X, \rho](x,r)\Big| N(x,r) = k\Big] = \frac{k}{(N-1)\mu_\rho(x,r)} \tag{7.10}$$

for all $k \in \{0, \ldots, N-1\}$. Therefore,

$$\begin{aligned}
\mathbb{E}\Big[\widehat{\mathrm{vol}}[d_X, \rho](x,r)\Big] &= \sum_{k=0}^{N-1} \mathbb{E}\Big[\widehat{\mathrm{vol}}[d_X, \rho](x,r)\Big| N(x,r) = k\Big] \cdot \mathbb{P}[N(x,r) = k] \\
&= \frac{1}{(N-1)\mu_\rho(x,r)} \sum_{k=0}^{N-1} k \cdot \mathbb{P}[N(x,r) = k] \\
&= \frac{\mathbb{E}[N(x,r)]}{(N-1)\mu_\rho(x,r)} \\
&= \mathrm{vol}(B^M(x,r))\,.
\end{aligned}$$

□

Lemma 7.2.5. Let X be a point cloud that consists of N points that are sampled from the probability density function $\rho : M \to \mathbb{R}_+$. If M is compact, then there is a constant $A > 0$ that only depends on ρ and the Riemannian metric of M and satisfies

$$\mathrm{var}(\widehat{\mathrm{vol}}[d_X, \rho](x,r)) \leq Ar^n/N$$

for sufficiently large N, sufficiently small r, and all $x \in X$.

Proof. By Lemma 7.A.2,

$$\mathrm{var}(\widehat{\mathrm{vol}}[d_X, \rho](x,r)) = \frac{\mathrm{var}(1/\rho(z)) \cdot \mu_\rho(x,r) \cdot \mathrm{vol}(B^M(x,r))}{(N-1)} + \frac{\mathrm{var}N(x,r)}{(N-1)^2\mu_\rho(x,r)^2}\,, \tag{7.11}$$

where $z \in B^M(x,r)$ is a point chosen randomly from the probability density function $\psi(z)$ defined as in Eq. 7.5 and

$$\mathrm{var}(N(x,r)) = (N-1)\mu_\rho(x,r)\mathrm{vol}(B^M(x,r))(1 - \mu_\rho(x,r)\mathrm{vol}(B^M(x,r)))\,.$$

Now we bound $\mathrm{var}(1/\rho(z))$. Define

$$A(r) := \max_{x \in M,\, z \in B^M(x,r)} |\rho(z) - \rho(x)| \,. \tag{7.12}$$

The quantity $A(r)$ exists because M is compact and ρ is continuous. We note that $A(r) \to 0$ as $r \to 0$. For the remainder of the proof, we assume that r is sufficiently small such that $A(r) \leq \min(\rho)/2$. Because $h(\rho) = 1/\rho^2$ is convex and monotonically decreasing for $\rho > 0$, we have

$$\Big|\frac{1}{\rho(z)^2} - \frac{1}{\rho(x)^2}\Big| \leq |h'(\min\{\rho(z), \rho(x)\})| \cdot A(r) \leq \frac{2A(r)}{(\rho(x) - A(r))^3} \leq \frac{16A(r)}{\min(\rho)^3}$$

for all $z \in B^M(x,r)$. Therefore,

$$\Big|\mathbb{E}[1/\rho(z)^2] - 1/\rho(x)^2\Big| \leq \frac{16A(r)}{\min(\rho)^3} \,.$$

Similarly,

$$\Big|\frac{1}{\mu_\rho(x,r)^2} - \frac{1}{\rho(x)^2}\Big| \leq \frac{16A(r)}{\min(\rho)^3}$$

because $|\mu_\rho(x,r) - \rho(x)| \leq A(r)$. Therefore,

$$\begin{aligned}
\mathrm{var}(1/\rho(z)) &= \Big|\mathbb{E}\Big[\frac{1}{\rho(z)^2}\Big] - \frac{1}{\mu_\rho(x,r)^2}\Big| \\
&\leq \Big|\mathbb{E}\Big[\frac{1}{\rho(z)^2}\Big] - \frac{1}{\rho(x)^2}\Big| + \Big|\frac{1}{\mu_\rho(x,r)^2} - \frac{1}{\rho(x)^2}\Big| \\
&\leq \frac{32A(r)}{\min(\rho)^3} \,.
\end{aligned} \tag{7.13}$$

$$\mathrm{var}(1/\rho(z)) \cdot \frac{\mu_\rho(x,r) \cdot \mathrm{vol}(B^M(x,r))}{(N-1)} \leq \frac{32 \cdot A(r) \max(\rho) \cdot \mathrm{vol}(B^M(x,r))}{\min(\rho)^3 (N-1)}$$

for sufficiently small r such that $A(r) < \min(\rho)/2$. By Lemma 7.A.1, there is a constant $B' > 0$ such that $\mathrm{vol}(B^M(x,r)) \leq B' r^n$ for all x and sufficiently small r. Additionally, we have $A(r) < 1$ for sufficiently small r, so

$$\mathrm{var}(1/\rho(z)) \cdot \frac{\mu_\rho(x,r) \cdot \mathrm{vol}(B^M(x,r))}{(N-1)} \leq \frac{32B \cdot \max(\rho)}{\min(\rho)^3} \cdot \frac{r^n}{N} \tag{7.14}$$

for sufficiently large N, sufficiently small r, and some constant $B > 0$. Lastly, we bound $\frac{\mathrm{var}N(x,r)}{(N-1)^2\mu_\rho(x,r)^2}$. We have

$$\begin{aligned}\frac{\mathrm{var}N(x,r)}{(N-1)^2\mu_\rho(x,r)^2} &= \frac{\mathrm{vol}(B^M(x,r))(1-\mu_\rho(x,r)\mathrm{vol}(B^M(x,r))}{(N-1)\mu_\rho(x,r)} \\ &\leq \frac{\mathrm{vol}(B^M(x,r))}{(N-1)\mu_\rho(x,r)} \\ &\leq \frac{B}{\min(\rho)} \cdot \frac{r^n}{N}\,, \end{aligned} \tag{7.15}$$

where $B > 0$ is the same constant as earlier in the proof. Substituting Equations 7.14 and 7.15 into Equation 7.11 completes the proof. □

7.2.5 Fitting a quadratic curve

For radius $r > 0$, let $y(x,r)$ and $\hat{y}[d_X, \hat{\rho}, \hat{n}](x,r)$ denote the actual and estimated ball-volume ratios, respectively, for a ball of radius r that is centered at a fixed $x \in X$. That is, we define

$$y(x,r) := \frac{\mathrm{vol}(B^M(x,r))}{v_n r^n}\,, \tag{7.16}$$

$$\hat{y}[d_X, \hat{\rho}, \hat{n}](x,r) := \frac{\widehat{\mathrm{vol}}[d_X, \hat{\rho}](x,r)}{v_{\hat{n}} r^{\hat{n}}}\,, \tag{7.17}$$

where $\widehat{\mathrm{vol}}[d_X, \hat{\rho}](x,r)$ is defined as in Eq. 7.7. When $\hat{\rho}(x) = \rho(x)$ for all $x \in X$, we write $\hat{y}[d_X, \rho, \hat{n}](x,r)$. When d_X, $\hat{\rho}$, or $\hat{n}$ are clear from context, we omit them from our notation.

Let $r_{\min}$ and $r_{\max}$, respectively, be the minimum and maximum ball radius that we consider, where $0 \leq r_{\min} < r_{\max}$. These are hyperparameters that must be set by a user. Let $r_0 := r_{\min} < r_1 < \cdots < r_m := r_{\max}$ be a monotonically increasing sequence, which is also set by a user. These are the radius values at which we estimate geodesic ball volumes by empirically approximating the maximum-likelihood estimator, as in Section 7.2.4. We allow any choice of sequence $\{r_i\}_{j=1}^m$, although we study only two possible choices in this chapter:

1. **Equal spacing:** The sequence is evenly spaced with spacing Δr. This is the choice that we make in Theorems 7.3.1 and 7.4.3.

2. **Nearest-neighbor distance:** In our numerical experiments, we allow $\{r_j\}$ to depend on x and set r_j to beequal to the distance from x to its jth nearest neighbor.

We define $C(x)$ to be the coefficient such that $1+C(x)r^2$ is the "best-fit" quadratic curve to the curve $y(x,r)$for $r \in [r_{\min}, r_{\max}]$. More precisely, we define

$$C(x) := \arg\min_{c\in\mathbb{R}} \left\| y(x,r) - (1+cr^2) \right\|_{L^2([r_{\min}, r_{\max}])} .$$

One can show that

$$C(x) = \frac{\int_{r_{\min}}^{r_{\max}} r^2 [y(x,r) - 1] dr}{\frac{1}{5}(r_{\max}^5 - r_{\min}^5)} .$$

We define

$$\hat{C}[d_X, \hat{\rho}, \hat{n}](x) := \frac{\sum_{i=1}^m r_i^2 (\hat{y}[d_X, \hat{\rho}, \hat{n}](x, r_i) - 1)(r_i - r_{i-1})}{\frac{1}{5}(r_{\max}^5 - r_{\min}^5)} \tag{7.18}$$

to be an estimate of $C(x)$. We omit d_X, $\hat{\rho}$, and $\hat{n}$ from our notation when they are clear from context.

7.2.6 Our scalar curvature estimate

Putting together Sections 7.2.1–7.2.5, we estimate scalar curvature.

Definition 7.2.6. Let d_X be a distsance matrix, let $\hat{\rho}$ be a density estimator, and let $\hat{n}$ be a dimension estimate. Given hyperparameters $r_{\min} \geq 0$ (the minimum ball radius that we consider), $r_{\max} > r_{\min}$ (the maximum ball radius that we consider), and $\{r_j\}_{j=0}^m$ (the sequence of ball radii that we consider, where $r_0 = r_{\min}$ and $r_m = r_{\max}$), our estimate of the scalar curvature at x is

$$\hat{S}[d_X, \hat{\rho}, \hat{n}](x) := -6(\hat{n}+2)\hat{C}[d_X, \hat{\rho}, \hat{n}](x) ,$$

where $\hat{C}[d_X, \hat{\rho}, \hat{n}](x)$ is defined in Equation 7.18.

When the distance matrix d_X, density estimator $\hat{\rho}$, and dimension estimate $\hat{n}$ are clear from context, we omit them and write $\hat{S}(x)$.

7.2.7 Computational complexity

In our numerical experiments (Section 7.5), we find that setting r_i equal to the distance to the ith nearest neighbor results in an estimate that is both accurate and computationally efficient. In this case,

$$\widehat{\mathrm{vol}}[d_X, \hat{\rho}](x, r_i) = \frac{1}{N-1} \sum_{j=1}^{m} \frac{1}{\hat{\rho}(z_j)} \,,$$

where $z_j \in X$ is the jth nearest neighbor of x. We precompute the pointwise density estimates $\hat{\rho}(z)$ for all $z \in X$. For every $x \in X$, we sort $\{d(x,z) \mid z \in X\}$ to compute its nearest neighbors $z_1, z_2, \ldots$ and its distance to those neighbors. Given these quantities, the set $\{\widehat{\mathrm{vol}}[d_X, \hat{\rho}](x, r_i)\}_{i=1}^{m}$ can be computed in $\mathcal{O}(m)$ time for any m because

$$\widehat{\mathrm{vol}}[d_X, \hat{\rho}](x, r_{i+1}) = \widehat{\mathrm{vol}}[d_X, \hat{\rho}](x, r_i) + \frac{1}{(N-1) \cdot \hat{\rho}(z_{i+1})} \,.$$

7.3 Stability

Most real-world data sets have errors or noise, so the given distances d_X likely differ from the true geodesic distances d. If the geodesic distances are estimated from a point cloud, errors are expected even if there is no noise in the data (i.e., the point cloud) itself. Density estimation introduces additional errors. Theorem 7.3.1 below says that our scalar curvature estimate $\hat{S}$ is stable with respect to errors in metric data and density estimation. This allows us to accurately estimate scalar curvature in real-world data or in synthetic point-cloud data in which distances are estimated.

Throughout this section, we consider a compact n-dimensional Riemannian manifold M with geodesic distance d and a sequence $\{X_k\}_{k=1}^{\infty}$ of point clouds that are sampled randomly from a probability density function $\rho : M \to \mathbb{R}_+$. We assume that $|X_k| \to \infty$ as $k \to \infty$. Let d_{X_k} denote the geodesic distance matrix for X_k. By a slight abuse of notation, let $d_{X_k}(x, y)$ denote the geodesic distance between points $x \in X_k$ and $y \in X_k$. We also consider sequences $\{r_{\min,\,k}\}_{k=1}^{\infty}$, $\{r_{\max,\,k}\}_{k=1}^{\infty}$, and $\{(\Delta r)_k\}_{k=1}^{\infty}$ of hyperparameter values. The kth radius sequence

that we consider is $\{r_{j,k}\}_{j=0}^{m_k}$, where $r_{j,k} := r_{\min,k} + j(\Delta r)_k$. When k is clear from context, we omit it and write r_j instead of $r_{j,k}$. We require that

1. $0 < r_{\min,k} < r_{\max,k}$ for all k,
2. the number $m_k := (r_{\max,k} - r_{\min,k})/(\Delta r)_k$ of radial steps is a positive integer for all k, and
3. $r_{\min,k} \to 0$, $r_{\max,k} \to 0$, and $(\Delta r)_k \to 0$ as $k \to \infty$.

Theorem 7.3.1 (Stability)**.** *For each k, suppose that $\widehat{d_{X_k}}$ is a metric on X_k such that*

$$\delta_k := \max_{x,x' \in X_k} |\widehat{d_{X_k}}(x,x') - d(x,x')| \to 0 \qquad \text{as } k \to \infty\,.$$

Suppose that $\hat{\rho}$ is a density estimator such that

$$\eta_k := \max_{x \in X_k} \left| \hat{\rho}[\widehat{d_{X_k}}](x) - \rho(x) \right| \to 0 \qquad \text{as } k \to \infty\,,$$

and suppose that $\hat{n}[\widehat{d_{X_k}}] = \hat{n}[d_{X_k}] = n$ for sufficiently large k. If the hyperparameter-value sequences satisfy

1. $\max_j \frac{A(2r_j)}{r_j^n r_{\max,k}^2} \to 0$ *as* $k \to \infty$,
2. $\eta_k/(r_{\min,k} + (\Delta r)_k)^{n+2/3} \to 0$ *as* $k \to \infty$,
3. $r_{\min,k} + (\Delta r)_k > \delta_k$ *for sufficiently large* k,
4. $|X_k|(\Delta r)_k(r_{\min,k} + (\Delta r)_k - \delta_k)^n \to \infty$ *as* $k \to \infty$,
5. $r_{\min,k}/r_{\max,k}^3 \to 0$ *as* $k \to \infty$,
6. $((\Delta r)_k + \delta_k)/r_{\max,k}^3 \to 0$ *as* $k \to \infty$, *and*
7. $((\Delta r)_k + \delta_k)/[(r_{\min,k} + (\Delta r)_k - \delta_k)^{n+1} r_{\max,k}^2] \to 0$ *as* $k \to \infty$,

then $|\hat{S}[\widehat{d_{X_k}}, \hat{\rho}, \hat{n}](x_k) - \hat{S}[d_{X_k}, \rho, \hat{n}](x_k)| \to 0$ in probability as $k \to \infty$, where $\{x_k\}$ is any sequence of points such that $x_k \in X_k$.

Remark 7.3.2. The conditions above on the hyperparameter-value sequences are complex. The following is a set of simpler conditions that collectively imply the conditions of Theorem 7.3.1:

1. $\frac{A(r)}{r^{n+2}} \to 0$ as $r \to 0$,
2. $\eta_k / r_{\min,k}^{n+2/3} \to 0$ as $k \to \infty$,
3. $\delta_k = \mathcal{O}(r_{\min,k}^{n+2})$ as $k \to \infty$,
4. $|X_k|(\Delta r)_k r_{\min,k}^n \to \infty$ as $k \to \infty$,
5. $(r_{\min,k})/r_{\max,k}^3 \to 0$ as $k \to \infty$, and
6. $(\Delta r)_k / r_{\min,k}^{n+5/3} \to 0$ as $k \to \infty$.

Proof of Theorem 7.3.1. For any $x \in X_k$, we have

$$\left|\hat{S}[\widehat{d_{X_k}}, \hat{\rho}](x) - \hat{S}[d_{X_k}, \rho](x)\right| = 6(n+2)\left|\hat{C}[\widehat{d_{X_k}}, \hat{\rho}](x) - \hat{C}[d_{X_k}, \rho](x)\right|.$$

The theorem follows from Lemma 7.A.5, which shows that

$$|\hat{C}[\widehat{d_{X_k}}, \hat{\rho}](x_k) - \hat{C}[d_{X_k}, \rho](x_k)| \to 0$$

in probability as $k \to \infty$. □

7.4 Convergence

Informally, what we show in Theorem 7.4.3 is that (1) as the number of points increases, (2) as our given metric data becomes more accurate, and (3) as our density estimations become more accurate, our scalar curvature estimate $\hat{S}(x)$ converges to the true scalar curvature $S(x)$. Throughout this section, the symbols M, d, n, $\{X_k\}_{k=1}^{\infty}$, d_{X_k}, ρ, $\{r_{\min,k}\}_{k=1}^{\infty}$, $\{r_{\max,k}\}_{k=1}^{\infty}$, $\{(\Delta r)_k\}_{k=1}^{\infty}$, m_k and $\{r_j\}_{j=1}^{m_k}$ are defined as in Section 7.3.

Theorem 7.4.3 is an immediate consequence of Theorem 7.3.1 (stability) and Prop. 7.4.2, a proposition (see below) that states that if we are given perfect metric data and the exact density, then our scalar curvature estimate $\hat{S}(x)$ converges to $S(x)$ as the number of points increases. The challenge is that we must take $r_{\max,k} \to 0$ for Equation 7.1 to hold, but (as we show in Prop. 7.4.1 below) the mean squared error of the ball-ratio estimate $\hat{y}[d_X, \rho](x, r)$ grows as $\mathcal{O}(1/(Nr^n))$ as $r \to 0$, where N is the number of points in the point cloud.

Proposition 7.4.1. Let X be a point cloud that consists of N points that are drawn from the probability density function $\rho : M \to \mathbb{R}_+$. Then there is a constant $A > 0$ that only depends on ρ and the Riemannian metric of M such that

$$\mathrm{MSE}(\hat{y}[d_X, \rho](x, r)) = \mathrm{var}(\hat{y}[d_X, \rho](x, r)) \leq \frac{A}{Nr^n}$$

for sufficiently large N, sufficiently small r, and all $x \in X$.

Proof. By Lemma 7.2.4,

$$\mathrm{MSE}(\hat{y}[d_X, \rho](x, r)) = \mathrm{var}(\hat{y}[d_X, \rho](x, r))\,.$$

By Lemma 7.2.5, there is a constant $A' > 0$ such that

$$\mathrm{var}(\widehat{\mathrm{vol}}[d_X, \hat{\rho}](x, r)) \leq \frac{A' r^n}{N}$$

for sufficiently large N, sufficiently small r, and all $x \in X$. Therefore,

$$\mathrm{var}(\hat{y}[d_X, \rho](x, r)) = \frac{\mathrm{var}(\hat{y}[d_X, \rho](x, r))}{v_n^2 r^{2n}} \leq \frac{A'}{v_n^2 N r^n}$$

for sufficiently large N, sufficiently small r, and all $x \in X$. □

Proposition 7.4.2. Suppose that $\hat{n}[d_{X_k}] = n$ for sufficiently large k. If the hyperparameter-value sequences satisfy

1. $(\Delta r)_k / r_{\max,k}^3 \to 0$ as $k \to \infty$,

2. $|X_k|(r_{\min,k} + (\Delta r)_k)^n \to \infty$ as $k \to \infty$, and

3. $r_{\min,k}/r_{\max,k}^3 \to 0$ as $k \to \infty$,

then $|\hat{S}[d_{X_k}, \rho, \hat{n}](x_k) \to S(x_k)| \to 0$ as $k \to \infty$, where $\{x_k\}$ is any sequence of points such that $x_k \in X_k$.

Proof. Let x be any point in X_k. By Eq. 7.1, we have

$$C(x) = \frac{\int_{r_{\min,k}}^{r_{\max,k}} \left[-\frac{S(x)}{6(n+2)} r^4 + \mathcal{O}(r^6) \right] dr}{\frac{1}{5}(r_{\max,k}^5 - r_{\min,k}^5)} = -\frac{S(x)}{6(n+2)} + \mathcal{O}(r_{\max,k}^2) .$$

The absolute difference $|\hat{S}[d_{X_k}, \rho](x) - S(x)|$ is

$$|\hat{S}[d_{X_k}, \rho](x) - S(x)| = 6(n+2)|\hat{C}[d_{X_k}, \rho](x) - C(x)| + \mathcal{O}(r_{\max,k}^2) .$$

Applying Lemma 7.A.6 yields the desired result. □

Theorem 7.4.3. *For each k, suppose that $\widehat{d_{X_k}}$ is a metric on X_k such that*

$$\delta_k := \max_{x,x' \in X_k} |\widehat{d_{X_k}}(x, x') - d(x, x')| \to 0 \qquad \text{as } k \to \infty .$$

Suppose that $\hat{\rho}$ is a density estimator such that

$$\eta_k := \max_{x \in X_k} \left| \hat{\rho}[\widehat{d_{X_k}}](x) - \rho(x) \right| \to 0 \qquad \text{as } k \to \infty .$$

Suppose that $\hat{n}[\widehat{d_{X_k}}] = \hat{n}[d_{X_k}] = n$ for sufficiently large k. If the hyperparameter-value sequences satisfy

1. $\max_j \frac{A(2r_j)}{r_j^n r_{\max,k}^2} \to 0$ *as* $k \to \infty$,

2. $\eta_k/(r_{\min,k} + (\Delta r)_k)^{n+2/3} \to 0$ *as* $k \to \infty$,

3. $r_{\min,k} + (\Delta r)_k > \delta_k$ *for sufficiently large* k,

4. $|X_k|(\Delta r)_k(r_{\min,k} + (\Delta r)_k - \delta_k)^n \to \infty$ *as* $k \to \infty$,

5. $r_{\min,k}/r^3_{\max,k} \to 0$ *as* $k \to \infty$,

6. $((\Delta r)_k + \delta_k)/r^3_{\max,k} \to 0$ *as* $k \to \infty$, *and*

7. $((\Delta r)_k + \delta_k)/[(r_{\min,k} + (\Delta r)_k - \delta_k)^{n+1} r^2_{\max,k}] \to 0$ *as* $k \to \infty$

then $|\hat{S}[\widehat{d_{X_k}}, \hat{\rho}, \hat{n}](x_k) - S(x_k)| \to 0$ *in probability as* $k \to \infty$, *where* $\{x_k\}$ *is any sequence of points such that* $x_k \in X_k$.

Remark 7.4.4. The simpler set of conditions from Remark 7.3.2 collectively implies the conditions of Theorem 7.4.3.

Proof of Theorem 7.4.3. The theorem follows from Theorem 7.3.1 and Prop. 7.4.2. □

7.5 Numerical Experiments

7.5.1 Data sets

We generate synthetic data by sampling uniformly at random from manifolds with known scalar curvature.

First, we sample $N = 10^4$ points each from three constant-curvature surfaces:

1. A disk in the Euclidean plane with radius 2. The scalar curvature is $S(x) \equiv 0$.

2. A unit 2-sphere. The scalar curvature is $S(x) \equiv 2$.

3. A disk in the hyperbolic plane with hyperbolic radius 2. The scalar curvature is $S(x) \equiv -2$.

For the last of these, we use the Poincaré disk model. Notably, the points that we sample from the hyperbolic plane are not embedded in Euclidean space, which means that it is not possible to use the scalar-curvature estimation method of [SWH21]. To avoid boundary

effects, we only estimate curvature at points within the unit disk in the Euclidean sample and within hyperbolic radius 1 in the hyperbolic sample. Additionally, we sample point clouds from S^2 with noise. For $\sigma \in \{.001, .003, .01, .03\}$, we sample $N = 10^4$ points from S^2 and add isotropic Gaussian noise with standard deviation σ.

Next, we sample point clouds from several other manifolds. We sample $N = 10^4$ points each from the higher-dimensional unit spheres S^n for $n = 3, 5$, and 7. Lastly, we sample one point cloud each from two surfaces with non-constant scalar curvature:

1. A 2-torus. We sample $N = 10^4$ points from a 2-torus with parameters $r = 1$, $R = 2$.

2. A one-sheet hyperboloid. The points $(x, y, z) \in \mathbb{R}^3$ are given by the equations

$$\begin{aligned} x &= 2\sqrt{1+u^2}\cos(\theta)\,, \\ y &= 2\sqrt{1+u^2}\sin(\theta)\,, \\ z &= u \end{aligned}$$

   for $u \in \mathbb{R}$ and $\theta \in [0, 2\pi)$. We sample points uniformly at random from the subset of the hyperboloid such that $|z| \leq 2$ until we have $N = 10^4$ points within the subset such that $|z| \leq 1$. To avoid boundary effects, we only estimate curvature at points on the hyperboloid such that $|z| \leq 1$.

7.5.2 Dimension estimation

To estimate dimension, we use the maximum-likelihood method of Levina and Bickel [LB04]. Our estimate of the dimension of a point cloud X is the nearest integer $\hat{n}$ to

$$\frac{1}{k_2 - k_1 + 1} \sum_{k=k_1}^{k_2} \hat{n}_k\,,$$

where k_1 and k_2 are hyperparameters and

$$\hat{n}_k := \frac{1}{N}\sum_{i=1}^{N} \hat{n}_k(x_i)\,,$$
$$\hat{n}_k(x_i) := \Big[\frac{1}{k-1}\sum_{j=1}^{k-1} \log\Big(\frac{T_k(x_i)}{T_j(x_i)}\Big)\Big]^{-1},$$

where $T_j(x_i)$ is the distance from x_i to its jth nearest neighbor in X. For all data sets, we set $k_1 = 20$ and calculate $\hat{n}$ for $k_2 \in \{30, \ldots, 100\}$. We obtain $\hat{n} = n$, where n is the ground-truth dimension, for all data sets and all choices of k_2.

We make one modification to [LB04], which is that instead of using Euclidean distance to measure distances to nearest neighbors, as was done in [LB04], we use geodesic distance.[2] This choice reduces overall computation time because computing geodesic nearest-neighbor distances is also part of our scalar-curvature estimation pipeline. In addition, using geodesic distance improves the accuracy of the approximations that were made in [LB04] and allows us to estimate the dimension of our Poincaré-disk data, which is not embedded in Euclidean space.

7.5.3 Density estimation

We use kernel density estimation to obtain pointwise estimates of density, using the dimension estimates obtained in Section 7.5.2. We test two choices of kernel: (1) a Gaussian kernel because it is a very common choice for density estimation and (2) a biweight kernel because it is compactly supported. As input, the kernel function takes geodesic distances (either exact or estimated), rather than Euclidean distances.

[2]In cases where we can calculate both exact and estimated geodesic distances, we use both; otherwise, we use whichever is available. For S^2, S^3, S^5, S^7, and the Euclidean disk, we possess both exact and estimated geodesic distances. For the Poincaré disk, we have only exact geodesic distances. For all other data sets, we have only estimated geodesic distances.

7.5.4 Geodesic-distance estimation

On the spheres, the Euclidean disk, the torus, and the hyperboloid, we estimate pairwise geodesic distances using the method of Tenenbaum et al. [TdL00, BSL00]. For each point cloud, we construct the k-nearest neighbor graph G with $k = 20$ for $n = 2$, with $k = 50$ for $n = 3$, with $k = 100$ for $n = 5$, and with $k = 200$ for $n = 7$. Edge weights are Euclidean distances. Our estimation of the geodesic distance between points x_1 and x_2 is the length of a shortest path in G.

7.5.5 Hyperparameter choices

Our method requires a choice of minimum ball radius $r_{\min}$, maximum ball radius $r_{\max}$, and radius sequence $\{r_j\}_{j=0}^m$ such that $r_0 = r_{\min}$ and $r_m = r_{\max}$. For a given point x in a data set, we set r_i equal to the distance from x to its ith nearest neighbor (as measured by the given distance matrix d_X), for the subset of neighbors such that $r_{\min} \leq r_i \leq r_{\max}$. We set $r_{\min} = 0$ for all data sets.

Our choice of $r_{\max}$ differs across data sets because the scales and sampling densities are different in different data sets. For the spheres (including the point clouds with noise), we set $r_{\max} = \pi/2$. For the Euclidean and Poincaré disks, we set $r_{\max} = 1$. For the torus, we set $r_{\max} = \pi$. For the hyperboloid, we set $r_{\max} = 2$. These values were chosen to minimize the amount of noise in our curvature estimation results and to ensure that our geodesic balls $B^M(x, r)$ do not intersect the boundary of the manifold M.

7.5.6 Results

First, we apply our method to our constant-curvature data sets. For the two surfaces that are embedded into Euclidean space (S^2 and the Euclidean disk), we test our method in two different ways. First, we use the exact geodesic distances for our distance matrix. Second, we estimate geodesic distances from the point clouds. In Figure 7.5.1, we show our results.

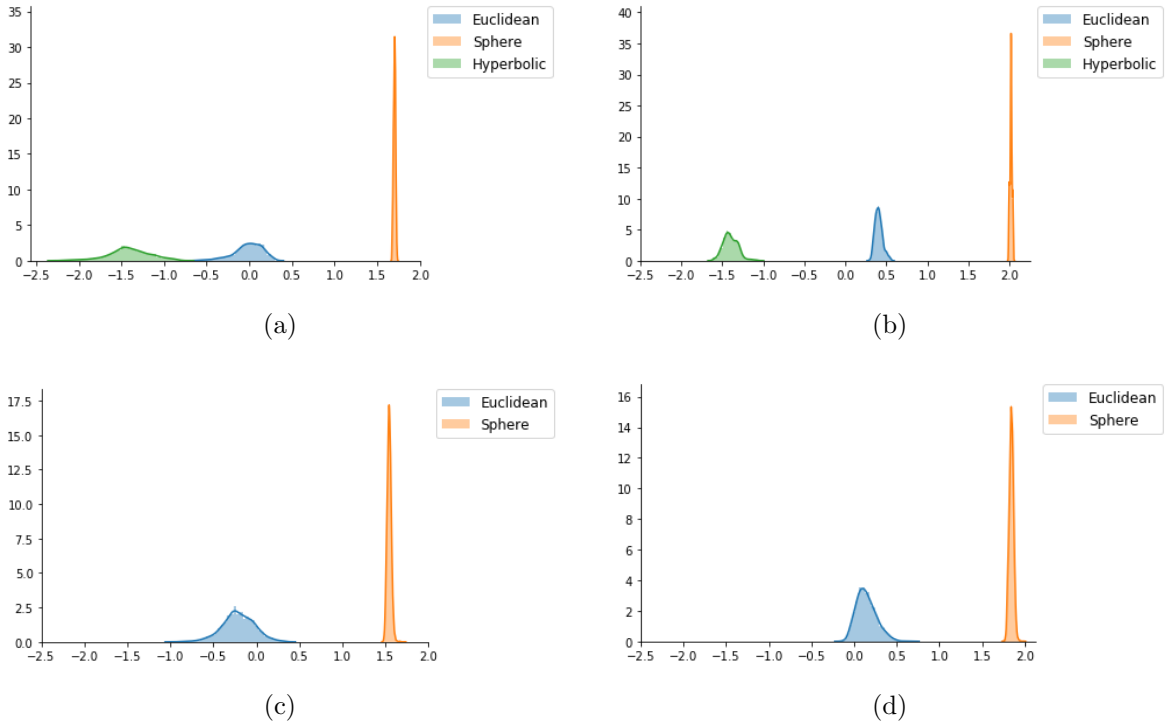
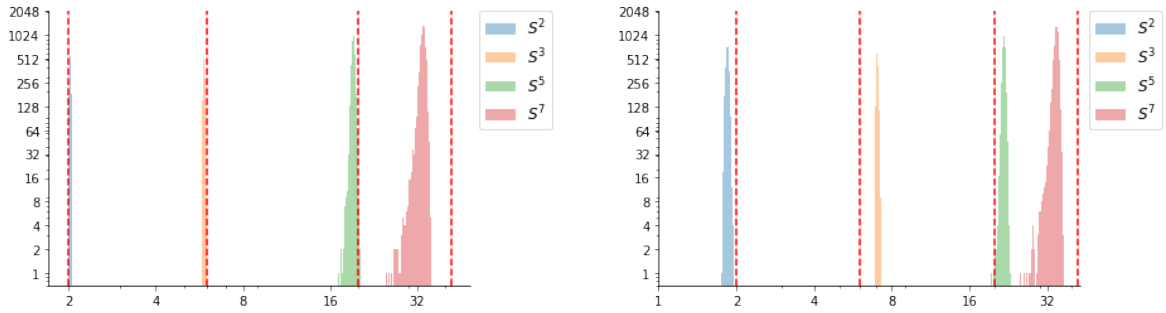


Figure 7.5.1: Histograms for our scalar-curvature estimates on three surfaces of constant curvature, given (A–B) exact geodesic distances and (B–C) point clouds, from which geodesic distances were estimated. In (A) and (C), we use a Gaussian kernel to estimate density, and in (B) and (D), we use a biweight kernel to estimate density. The ground-truth scalar curvatures values are -2 in the hyperbolic disk, 0 in the Euclidean disk, and 2 on the sphere.

We next test our method on the point clouds that are sampled from higher-dimensional spheres. Again, we test our method once in which we input exact geodesic distances and a second time in which we estimate geodesic distance from the point clouds. In early experiments, we found that on the highest-dimensional spheres ($n \geq 5$), using a biweight kernel to estimate density led to significantly better performance than using a Gaussian kernel, so we use a biweight kernel for density estimation. In Figure 7.5.2, we show our results. Unexpectedly, we find in Figure 7.5.2(A) that scalar curvature is systematically underestimated (although still reasonably accurate) when we start with the exact geodesic distances. In both experiments, the accuracy of our estimates decreases as the dimension n increases, but the performance is comparable to that in [SWH21]. The main reason that scalar curvature is more difficult for us to estimate in higher dimensions is that the mean squared error in our ball-ratio estimates increases exponentially in n (see Prop. 7.4.1). Another reason is that the accuracy of geodesic-distance estimation decreases as n increases and N stays constant. Typically, the number of points N must scale exponentially with n to maintain the same "resolution" of the manifold, so it is unsurprising that our scalar curvature estimates become less accurate as n increases for fixed N.

To test our method on manifolds with non-constant scalar curvature, we apply our scalar-curvature estimator to our torus and hyperboloid data sets. On both surfaces, we find that using a Gaussian kernel for density estimation yields more accurate curvature estimates, so we use a Gaussian kernel. We show our results in Figure 7.5.3. On the torus, our estimator correctly distinguishes between regions of positive, negative, and zero scalar curvature. The estimates are accurate except near $\theta = \pi$, where scalar curvature is minimized. On the hyperboloid, our estimator correctly identifies the fact that scalar curvature is minimized (and negative) near $z = 0$ and increases as z increases. As in the torus, the estimates are accurate except near $z = 0$, where scalar curvature is minimized.

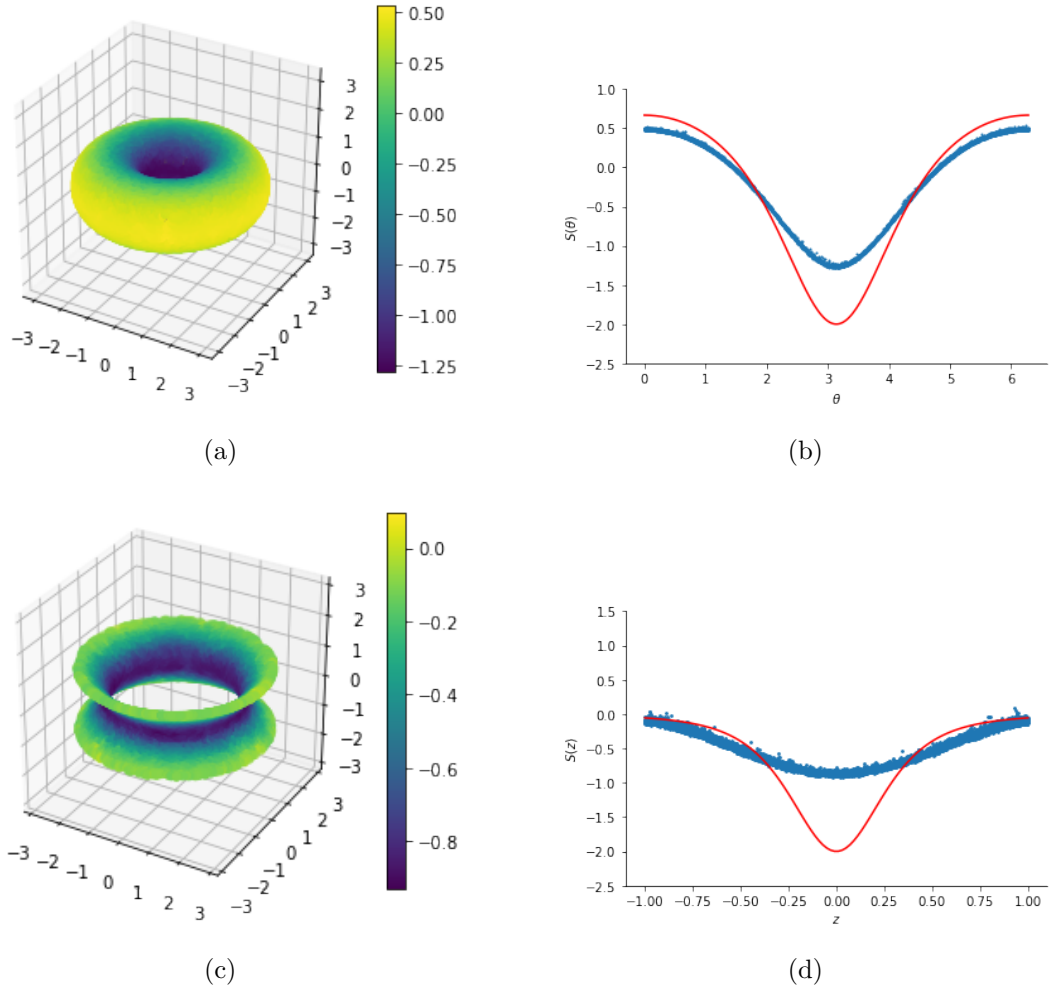
We investigate the stability of our estimator by estimating curvature on our noisy-sphere data sets. We show our results in Figure 7.5.4. In Figures 7.5.4(A) and (B), we show



(a) (b)

Figure 7.5.2: Histograms for our scalar-curvature estimates on S^n (for $n = 2, 3, 5, 7$) using (A) exact geodesic distances and (B) point clouds, from which geodesic distances were estimated. In (A) and (B), the histograms are plotted on a log–log scale. The ground-truth scalar curvature, which is indicated by the red dashed lines, is $S(x) \equiv n(n-1)$ for each n and all $x \in S^n$. To improve visualization, we removed about 10 of the curvature estimates on S^7 that were extreme outliers.

our results when we use Gaussian and biweight kernels, respectively, for density estimation and we input the estimated geodesic distances to the kernel. At the highest noise level (standard deviation $\sigma = .03$), our scalar curvature estimates have the wrong sign when we use a biweight kernel, but all other curvature estimates have the correct sign. In Figures 7.5.4(C) and (D), we test our estimator by inputting Euclidean distances to the kernel for density estimation. We find that performance is significantly improved, especially at the highest noise level ($\sigma = .03$). This suggests that if a point cloud has a high noise level, then one should input the Euclidean distances to the kernel instead of inputting the estimated geodesic distances, which may not be accurate enough.



(a) (b)

(c) (d)

Figure 7.5.3: (A) Scalar-curvature estimation on a torus. (B) Scalar curvature on the torus as a function of angle θ. In red, we show the exact scalar curvature values; in blue, we show the estimated scalar curvature values. (C) Scalar-curvature estimation on a one-sheet hyperboloid. (D) Scalar curvature on the hyperboloid as a function of the z coordinate. In red, we show the exact scalar curvature values; in blue, we show the estimated scalar curvature values.

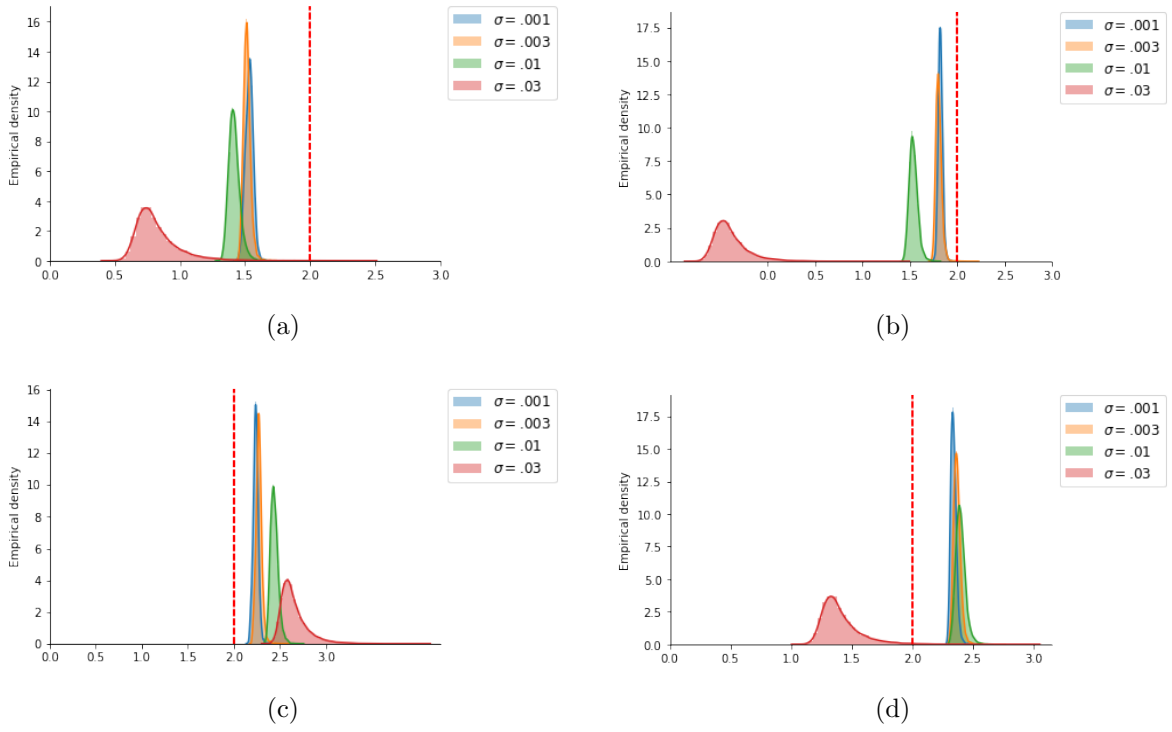


Figure 7.5.4: Scalar-curvature estimation on S^2 with isotropic Gaussian noise (standard deviation σ) added to the point cloud. (A) We use a Gaussian kernel for density estimation. The kernel takes the estimated geodesic distances as input. (B) We use a biweight kernel that takes estimated geodesic distances as input. (C) We use a Gaussian kernel that takes Euclidean distances as input. (D) We use a biweight kernel that takes Euclidean distances as input.

7.6 Conclusions

In this chapter, we developed a new method to estimate scalar curvature in discrete data. The only information that our approach requires is the set of pairwise distances between the points. By contrast, prior methods were limited to surfaces in $\mathbb{R}^3$ or to point clouds embedded in Euclidean space. Because our method depends only on metric data, one can use it to estimate curvature not only in point-cloud data (from which geodesic distances can be estimated using the approach in [TdL00, BSL00], for example), but also at vertices in a graph that is equipped with the shortest-path metric or at points in any finite metric space. We proved that under suitable conditions, our estimator is stable (Theorem 7.3.1) and that it converges to the ground-truth scalar curvature (Theorem 7.4.3).

We validated our method on several synthetic data sets in Section 7.5. Notably, our experiments included a data set (a point cloud that is sampled from the Poincaré disk) for which we possessed only the pairwise exact geodesic distances, not an embedding of the points in Euclidean space. Our experiments on point-cloud data embedded in Euclidean space are equivalent to experiments on *geometric graphs*, which are graphs in which vertices are sampled from a manifold, edges connect nearby points, and edge weights are given by distances. This is because we estimated geodesic distance in our point clouds by constructing a nearest-neighbor graph (which is a type of geometric graph) and computing shortest-path lengths. Therefore, our method for scalar-curvature estimation on a point cloud is equivalent to scalar-curvature estimation on the nearest-neighbor graph equipped with the shortest-path metric. Our experiments show that one can achieve reasonable accuracy even without having or using a Euclidean embedding of the data.

The primary limitation of our estimator is that it can be inaccurate on regions with non-constant curvature, especially near points on a manifold where a local extremum in the curvature is attained. (For example, see our experiments on the torus and hyperboloid in Section 7.5.) The reason is that when the radius r is small, we cannot reliably estimate the

ratio between $\text{vol}(B^M(x,r))$ and the volume of a Euclidean ball of radius r (see Prop. 7.4.1). We addressed this by using a relatively high $r_{\max}$ parameter, which controls the maximum ball radius that we consider. However, requiring r to be relatively large has the drawback that we are unable to detect local variation in scalar curvature; we are effectively smoothing out the curvature. In future work, we plan to investigate strategies to increase the accuracy of our method on manifolds with non-constant scalar curvature.

We expect that our scalar-curvature estimator will improve with improvements in state-of-the-art methods for density and geodesic-distance estimation on manifolds. Our method involves density estimation on a manifold as an intermediary step, and it also requires geodesic-distance estimation when we are given a point cloud embedded in Euclidean space instead of a distance matrix d_X. There are several other methods for geodesic-distance estimation that we did not use in our experiments; see [AHH19, LD19], for example. Improvements to the intermediary steps of our pipeline will lead to better performance of our scalar-curvature estimator.

It would also be interesting to incorporate machine learning into our curvature-estimation pipeline. For example, at each point, we estimate a sequence of ball-volume ratios (see Eq. 7.17); this is a vector that one can feed into a neural network, rather than using the method in Section 7.2.5 for estimating a quadratic coefficient. One could also use a graph neural network in which the graph is the nearest-neighbor graph for the data set and the initial node features are the vectors of ball-volume ratio estimates. Using machine learning would allow one to sidestep the choices of hyperparameters (the maximum ball radius $r_{\max}$, the minimum ball radius $r_{\min}$, and the radius sequence $\{r_j\}$), although those decisions would be replaced by different hyperparameter choices (e.g., a choice of learning rate). However, our current approach has the advantage that it is highly interpretable. We have designed our method so that, at minimum, one can reliably trust that the scalar curvature sign is accurate—in many cases, the sign of the curvature is the qualitative information that matters most—and that our method will generalize to manifolds that are not present in the training data set.

APPENDIX

7.A Technical details of Chapter 7

Here we prove some technical lemmas for proving our stability theorem (Theorem 7.3.1) and convergence theorem (Theorem 7.4.3). The notation that we use is the same as in Sections 7.3 and 7.4.

Lemma 7.A.1. If M is compact, then there are positive constants $B^{(1)}$ and $B^{(2)}$ such that

$$B^{(1)}r^n \leq \mathrm{vol}(B^M(x,r)) \leq B^{(2)}r^n \tag{7.19}$$

for sufficiently small r and all x in M.

Proof. By Equation 7.1, there are positive constants $B_x^{(1)}$, $B_x^{(2)}$, and r_x for each $x \in M$ such that

$$B_x^{(1)}r^n \leq \mathrm{vol}(B^M(x,r)) \leq B_x^{(2)}r^n \qquad \text{for } r < r_x\,.$$

Because the Riemannian metric g is smooth, the quantities r'_x, $B_x^{(1)}$, and $B_x^{(2)}$ can be chosen for each $x \in M$ such that the functions $x \mapsto r'_x$, $x \mapsto B_x^{(1)}$, and $x \mapsto B_x^{(2)}$ are continuous. If M is compact, then $B^{(i)} := \max_{x\in M} B_x^{(i)}$ and $r_* := \min_{x\in M} r_x$ exist, so Eq. 7.19 holds for $r < r_*$ and all x in M. □

Lemma 7.A.2. Let X be a point cloud that consists of N points drawn from probability density function $\rho : M \to \mathbb{R}_+$. Then

$$\mathrm{var}(\widehat{\mathrm{vol}}[d_X,\rho](x,r)) = \frac{\mathrm{var}(1/\rho(z)) \cdot \mu_\rho(x,r) \cdot \mathrm{vol}(B^M(x,r))}{N-1} + \frac{\mathrm{var}N(x,r)}{(N-1)^2\mu_\rho(x,r)^2}\,,$$

where $z \in B^M(x,r)$ is a point chosen randomly from the probability density function $\psi(z)$ defined in Eq. 7.5 and

$$\mathrm{var}(N(x,r)) = (N-1)\mu_\rho(x,r)\mathrm{vol}(B^M(x,r))(1-\mu_\rho(x,r)\mathrm{vol}(B^M(x,r)))\,.$$

Proof. By Lemma 7.2.4,

$$\operatorname{var}(\widehat{\operatorname{vol}}[d_X, \rho](x, r)) = \mathbb{E}\Big[(\widehat{\operatorname{vol}}[d_X, \rho](x, r) - \operatorname{vol}(B^M(x, r)))^2\Big].$$

By Equation 7.10,

$$\begin{aligned}
&\mathbb{E}\Big[(\widehat{\operatorname{vol}}[d_X, \rho](x, r) - \operatorname{vol}(B^M(x, r)))^2 \Big| N(x, r) = k\Big] \\
&\quad = \mathbb{E}[\widehat{\operatorname{vol}}[d_X, \rho](x, r)^2 \mid N(x, r) = k] - \frac{2k\operatorname{vol}(B^M(x, r))}{(N-1)\mu_\rho(x, r)} + \operatorname{vol}(B^M(x, r))^2 \qquad (7.20)
\end{aligned}$$

for all $k \in \{0, \ldots, N-1\}$. By Equation 7.8,

$$\mathbb{E}[\widehat{\operatorname{vol}}[d_X, \rho](x, r)^2 \mid N(x, r) = k] = \frac{k^2}{(N-1)^2} \cdot \mathbb{E}\Big[\Big(\frac{1}{k}\sum_{i=1}^{k} 1/\rho(z_i)\Big)^2\Big], \qquad (7.21)$$

where $\{z_i\}_{i=1}^k = B^M(x, r) \cap (X \setminus \{x\})$. If $k \geq 1$, the quantity $\frac{1}{k}\sum_{i=1}^k 1/\rho(z_i)$ is a sample mean. Therefore,

$$\mathbb{E}\Big[\Big(\frac{1}{k}\sum_{i=1}^{k} 1/\rho(z_i)\Big)^2\Big] = \frac{\operatorname{var}(1/\rho(z))}{k} + \mathbb{E}[1/\rho(z)]^2 = \frac{\operatorname{var}(1/\rho(z))}{k} + \frac{1}{\mu_\rho(x, r)^2}, \qquad (7.22)$$

where z is chosen from the probability density function $\psi(z)$ defined in Equation 7.5. The last equality follows by Equation 7.6. Substituting Equation 7.22 into Equation 7.21 and Equation 7.21 into 7.20, we obtain

$$\begin{aligned}
&\mathbb{E}\Big[(\widehat{\operatorname{vol}}[d_X, \rho](x, r) - \operatorname{vol}(B^M(x, r)))^2 \Big| N(x, r) = k\Big] \\
&= \Bigg[\frac{k}{(N-1)^2} \cdot \operatorname{var}\Big(\frac{1}{\rho(z)}\Big) \\
&\qquad + \frac{k^2 - 2k(N-1)\mu_\rho(x, r)\operatorname{vol}(B^M(x, r)) + (N-1)^2\mu_\rho(x, r)^2\operatorname{vol}(B^M(x, r))^2}{(N-1)^2\mu_\rho(x, r)^2}\Bigg] \qquad (7.23)
\end{aligned}$$

for $k \in \{1, \ldots, N-1\}$. When $k = 0$, Equation 7.23 holds because the right-hand side equals $\operatorname{vol}(B^M(x, r))$ and $\mathbb{E}[(\widehat{\operatorname{vol}}[d_X, \rho](x, r) - \operatorname{vol}(B^M(x, r)))^2 \mid N(x, r) = 0] = \operatorname{vol}(B^M(x, r))$.

To simplify the right-hand side of Equation 7.23, we observe that $N(x, r)$ is a binomial random variable with $N-1$ trials and success probability $\mu_\rho(x, r)\operatorname{vol}(B^M(x, r))$, so

$$\mathbb{E}[N(x, r)] = (N-1)\mu_\rho(x, r)\operatorname{vol}(B^M(x, r))$$

and

$$\begin{aligned}&\mathbb{E}[(N(x,r)-\mathbb{E}N(x,r))^2 \mid N(x,r)=k]\\&\quad=k^2-2k(N-1)\mu_\rho(x,r)\mathrm{vol}(B^M(x,r))+(N-1)^2\mu_\rho(x,r)^2\mathrm{vol}(B^M(x,r))\,.\end{aligned}$$

Therefore

$$\begin{aligned}&\mathbb{E}\Big[(\widehat{\mathrm{vol}}[d_X,\rho](x,r)-\mathrm{vol}(B^M(x,r)))^2\Big|N(x,r)=k\Big]\\&\quad=\frac{k}{(N-1)^2}\cdot\mathrm{var}(1/\rho(z))+\frac{1}{(N-1)^2\mu_\rho(x,r)^2}\cdot\mathbb{E}[(N(x,r)-\mathbb{E}N(x,r))^2\mid N(x,r)=k]\,.\end{aligned}$$

Putting it all together, we have that $\mathrm{var}(\widehat{\mathrm{vol}}[d_X,\rho](x,r))$ equals

$$\begin{aligned}&\mathbb{E}\Big[(\widehat{\mathrm{vol}}[d_X,\rho](x,r)-\mathrm{vol}(B^M(x,r)))^2\Big]\\&\quad=\sum_{k=0}^{N-1}\mathbb{E}\Big[(\widehat{\mathrm{vol}}[d_X,\rho](x,r)-\mathrm{vol}(B^M(x,r)))^2\Big|N(x,r)=k\Big]\mathbb{P}[N(x,r)=k]\\&\quad=\left(\frac{\mathrm{var}(1/\rho(z))}{(N-1)^2}\sum_{k=0}^{N-1}k\cdot\mathbb{P}[N(x,r)=k]\right.\\&\qquad\left.+\frac{1}{(N-1)^2\mu_\rho(x,r)^2}\sum_{k=0}^{N-1}\mathbb{E}[(N(x,r)-\mathbb{E}N(x,r))^2\mid N(x,r)=k]\mathbb{P}[N(x,r)=k]\right)\\&\quad=\frac{\mathrm{var}(1/\rho(z))}{(N-1)^2}\mathbb{E}[N(x,r)]+\frac{\mathrm{var}(N(x,r))}{(N-1)^2\mu_\rho(x,r)^2}\\&\quad=\frac{\mathrm{var}(1/\rho(z))\cdot\mu_\rho(x,r)\cdot\mathrm{vol}(B^M(x,r))}{(N-1)}+\frac{\mathrm{var}(N(x,r))}{(N-1)^2\mu_\rho(x,r)^2}\,,\end{aligned}$$

where

$$\mathrm{var}(N(x,r))=(N-1)\mu_\rho(x,r)\mathrm{vol}(B^M(x,r))(1-\mu_\rho(x,r)\mathrm{vol}(B^M(x,r)))$$

because $N(x,r)$ is a binomial random variable with parameters $N-1$ and $\mu_\rho(x,r)\mathrm{vol}(B^M(x,r))$.

□

Lemma 7.A.3. Assume that $\hat{n}[d_{X_k}]=n$ for sufficiently large k. Let $\{a_k\}_{k=1}^\infty$ and $\{b_k\}_{k=1}^\infty$ be sequences of positive real numbers such that

1. $0 < a_k < b_k$ for all k,

2. $a_k, b_k \to 0$ as $k \to \infty$, and

For each k, let R_k be a finite subset of $[a_k, b_k]$ such that $\frac{|R_k|}{|X_k| a_k^n} \to 0$ as $k \to 0$. Then

$$\max_{r \in R_k} |\hat{y}[d_{X_k}, \rho](x_k, r) - 1| \to 0$$

in probability as $k \to \infty$, where $\{x_k\}$ is any sequence of points such that $x_k \in X_k$.

Proof. Let $\epsilon > 0$. To simplify our notation, we denote $\hat{y}[d_{X_k}, \rho](x, r)$ by $\hat{y}(x, r)$. For any $x \in X_k$ and any $r \in [a_k, b_k]$,

$$\mathbb{P}\Big[|\hat{y}(x,r) - 1| > \epsilon\Big] \leq \mathbb{P}\Big[|\hat{y}(x,r) - y(x,r)| + |y(x,r) - 1| > \epsilon\Big]\,. \tag{7.24}$$

By Equation 7.1, there are constants $A > 0$ and $r_1 > 0$ such that

$$|y(x,r) - 1| \leq Ar^2 \qquad \text{for } r < r_1 \text{ and all } x \in M\,.$$

Let $r_2 = \min(\sqrt{\epsilon/(2A)}, r_1)$. If $r < r_2$, then $|y(x,r) - 1| < \frac{\epsilon}{2}$. For sufficiently large k, we have $b_k < r_2$, so by Equation 7.24,

$$\mathbb{P}\Big[|\hat{y}(x,r) - 1| > \epsilon\Big] \leq \mathbb{P}\Big[|\hat{y}(x,r) - y(x,r)| > \epsilon/2\Big] \tag{7.25}$$

for any $r \in [a_k, b_k]$ and for sufficiently large k. By Chebyshev's inequality,

$$\mathbb{P}\Big[|\hat{y}(x,r) - y(x,r)| > \epsilon/2\Big] \leq \frac{4\mathrm{var}(\hat{y}(x,r))}{\epsilon^2}\,. \tag{7.26}$$

By Proposition 7.4.1, there are positive constants B and $r_3 < r_2$ such that

$$\mathrm{var}(\hat{y}(x,r)) \leq \frac{B}{|X_k| r^n}$$

for sufficiently large k, all $r < r_3$, and any $x \in X_k$. Substituting into Equation 7.26 shows that

$$\mathbb{P}\Big[|\hat{y}(x,r) - y(x,r)| > \epsilon/2\Big] \leq \frac{4B}{\epsilon^2 |X_k| r^n}$$

for any $r < r_3$ and any $x \in X_k$. For sufficiently large k, we have $b_k < r_3$, so

$$\mathbb{P}\Big[|\hat{y}(x,r) - y(x,r)| > \epsilon/2\Big] \leq \frac{4B}{\epsilon^2 |X_k| a_k^n}$$

for any $r \in [a_k, b_k]$ and for sufficiently large k. Therefore,

$$\mathbb{P}\Big[\max_{r \in R_k} |\hat{y}(x,r) - y(x,r)| > \epsilon/2\Big] \leq \frac{4B|R_k|}{\epsilon^2 |X_k| a_k^n}$$

By hypothesis, the right-hand side approaches 0 as $k \to \infty$ because $\frac{|R_k|}{|X_k| a_k^n} \to 0$. Applying Equation 7.25 concludes the proof. □

Lemma 7.A.4 (Stability of $\hat{y}$)**.** For each k, suppose that $\widehat{d_{X_k}}$ is a metric on X_k such that

$$\delta_k := \max_{x,x' \in X_k} |\widehat{d_{X_k}}(x,x') - d(x,x')| \to 0 \qquad \text{as } k \to \infty$$

and $\hat{\rho}$ is a density estimator such that

$$\eta_k := \max_{x \in X_k} |\hat{\rho}[\widehat{d_{X_k}}](x) - \rho(x)| \to 0 \qquad \text{as } k \to \infty\,.$$

Suppose that $\hat{n}[\widehat{d_{X_k}}] = \hat{n}[d_{X_k}] = n$ for sufficiently large k. Additionally, suppose that the hyperparameter-value sequences satisfy the conditions:

1. $(r_{\min,\,k} + (\Delta r)_k)/r_{\max,\,k}^3 \to 0$ as $k \to \infty$.
2. $\eta_k/(r_{\min,\,k} + (\Delta r)_k)^{n+2/3} \to 0$ as $k \to \infty$.
3. $\max_j \frac{A(r_j)}{r_j^n r_{\max,\,k}^2} \to 0$ as $k \to \infty$, where $A(r)$ is defined as in Equation 7.12.
4. $r_{\min,\,k} + (\Delta r)_k - \delta_k > 0$ for sufficiently large k.
5. $\frac{\delta_k + (\Delta r)_k}{(r_{\min,\,k} + (\Delta r)_k - \delta_k)^{n+1} r_{\max,\,k}^2} \to 0$ as $k \to \infty$.

Define $\ell_k := \min\{\ell \in \mathbb{Z} \mid \ell(\Delta r)_k \geq \delta_k\}$. Then there is a sequence $\{\xi_k\}$ of nonnegative real numbers satisfying $\xi_k / r_{\max,\,k}^2 \to 0$ as $k \to \infty$ such that for any sequence $\{x_k\}_{k=1}^{\infty}$, where $x_k \in X_k$ for all k,

$$\hat{y}[d_{X_k}, \rho](x_k, r_{j-\ell_k}) - \xi_k \leq \hat{y}[\widehat{d_{X_k}}, \hat{\rho}](x_k, r_j) \leq \hat{y}[d_{X_k}, \rho](x_k, r_{j+\ell_k}) + \xi_k$$

for all $j \geq 2$ and

$$\hat{y}[d_{X_k}, \rho](x_k, r_1 - \delta_k) - \xi_k \leq \hat{y}[\widehat{d_{X_k}}, \hat{\rho}](x_k, r_1) \leq \hat{y}[d_{X_k}, \rho](x_k, r_{1+\ell_k}) + \xi_k$$

for $j = 1$.

Proof. For convenience, let $\hat{y}_k(x, r)$ denote $\hat{y}[\widehat{d_{X_k}}, \hat{\rho}](x, r)$ and let $\hat{y}(x, r)$ denote $\hat{y}[d_{X_k}, \rho](x, r)$. Define

$$\lambda_{j,k}^{+} := \ell_k(\Delta r)_k \,,$$

$$\lambda_{j,k}^{-} := \begin{cases} \ell_k(\Delta r)_k \,, & j \geq 2 \\ \delta_k \,, & j = 1 \end{cases}$$

for all j and k. Our goal is to compare $\hat{y}_k(x, r_j)$ to both $\hat{y}(x, r_j - \lambda_{j,k}^{-})$ and $\hat{y}(x, r_j + \lambda_{j,k}^{+})$. The "radial shift values" $\lambda_{j,k}^{\pm}$ are defined so that they satisfy

1. $r_j \pm \lambda_{j,k}^{\pm} > 0$ (all shifted radius values are positive) and
2. $\delta_k \leq \lambda_{j,k}^{\pm} \leq \delta_k + (\Delta r)_k$

for all j and sufficiently large k. For the remainder of the proof, we consider only k sufficiently large such that (1) holds. The key is that because $\lambda_{j,k}^{\pm} \geq \delta_k$ for all j and k,

$$N[d_{X_k}](x, r - \lambda_{j,k}^{-}) \leq N[\widehat{d_{X_k}}](x, r) \leq N[d_{X_k}](x, r + \lambda_{j,k}^{+}) \tag{7.27}$$

for all $x \in X_k$ and $r \geq 0$. We use Eq. 7.27 to compare $\widehat{\mathrm{vol}}[\widehat{d_{X_k}}, \hat{\rho}](x, r)$ and $\widehat{\mathrm{vol}}[d_{X_k}, \rho](x, r \pm \lambda_{j,k}^{\pm})$. First we quantify the error introduced by the error in density estimation. Observe that

$$\begin{aligned} \left|\widehat{\mathrm{vol}}[\widehat{d_{X_k}}, \hat{\rho}](x, r) - \widehat{\mathrm{vol}}[\widehat{d_{X_k}}, \rho](x, r)\right| &= \left|\frac{N[\widehat{d_{X_k}}](x, r)}{(|X_k| - 1)\widehat{\mu}_\rho[\hat{\rho}](x, r)} - \frac{N[\widehat{d_{X_k}}](x, r)}{(|X_k| - 1)\widehat{\mu}_\rho[\rho](x, r)}\right| \\ &\leq \left|\frac{1}{\widehat{\mu}_\rho[\hat{\rho}](x, r)} - \frac{1}{\widehat{\mu}_\rho[\rho](x, r)}\right|. \\ &\leq \left|\frac{1}{\widehat{\mu}_\rho[\hat{\rho}](x, r)} - \frac{1}{\rho(x)}\right| + \left|\frac{1}{\widehat{\mu}_\rho[\rho](x, r)} - \frac{1}{\rho(x)}\right| \end{aligned} \tag{7.28}$$

If $N[\widehat{d_{X_k}}](x,r) \geq 1$ and $d_{X_k}(x,z) \leq r$, then

$$\Big|\frac{1}{\rho(z)} - \frac{1}{\rho(x)}\Big| \leq \frac{A(r)}{\min(\rho)^2}$$

and

$$\begin{aligned}\Big|\frac{1}{\hat{\rho}(z)} - \frac{1}{\rho(x)}\Big| &\leq \Big|\frac{1}{\hat{\rho}(z)} - \frac{1}{\rho(z)}\Big| + \Big|\frac{1}{\rho(z)} - \frac{1}{\rho(x)}\Big| \\ &\leq \frac{\eta_k}{(\min(\rho) - \eta_k)^2} + \frac{A(r)}{\min(\rho)^2}\,.\end{aligned}$$

If $N[\widehat{d_{X_k}}](x,r) = 0$, then

$$\begin{aligned}&\Big|\frac{1}{\widehat{\mu}_\rho[\hat{\rho}](x,r)} - \frac{1}{\rho(x)}\Big| = \Big|\frac{1}{\hat{\rho}(x)} - \frac{1}{\rho(x)}\Big| \leq \frac{\eta_k}{(\min(\rho) - \eta_k)^2}\,, \\ &\Big|\frac{1}{\widehat{\mu}_\rho[\rho](x,r)} - \frac{1}{\rho(x)}\Big| = 0\,.\end{aligned}$$

Therefore,

$$\Big|\frac{1}{\widehat{\mu}_\rho[\hat{\rho}](x,r)} - \frac{1}{\rho(x)}\Big| + \Big|\frac{1}{\widehat{\mu}_\rho[\rho](x,r)} - \frac{1}{\rho(x)}\Big| \leq \frac{\eta_k}{(\min(\rho) - \eta_k)^2} + \frac{2A(r)}{\min(\rho)^2}\,.$$

By Eq. 7.28,

$$\begin{aligned}\widehat{\mathrm{vol}}[\widehat{d_{X_k}}, \rho](x, r_j) - \Big(\frac{\eta_k}{(\min(\rho) - \eta_k)^2} + \frac{2A(r_j)}{\min(\rho)^2}\Big) &\leq \widehat{\mathrm{vol}}[\widehat{d_{X_k}}, \hat{\rho}](x, r_j) \\ \leq \widehat{\mathrm{vol}}[\widehat{d_{X_k}}, \rho](x, r_j) + \frac{\eta_k}{(\min(\rho) - \eta_k)^2} + \frac{2A(r_j)}{\min(\rho)^2}\,.&\end{aligned} \tag{7.29}$$

Together, Equations 7.27 and 7.29 show that

$$\begin{aligned}\frac{N[d_{X_k}](x, r_j - \lambda^-_{j,k})}{(|X_k| - 1)\widehat{\mu}_\rho[\rho](x, r_j)} - \Big(\frac{\eta_k}{(\min(\rho) - \eta_k)^2} + \frac{2A(r_j)}{\min(\rho)^2}\Big) &\leq \widehat{\mathrm{vol}}[\widehat{d_{X_k}}, \hat{\rho}](x, r_j) \\ \leq \frac{N[d_{X_k}](x, r_j + \lambda^+_{j,k})}{(|X_k| - 1)\widehat{\mu}_\rho[\rho](x, r_j)} + \Big(\frac{\eta_k}{(\min(\rho) - \eta_k)^2} + \frac{2A(r_j)}{\min(\rho)^2}\Big)\,.&\end{aligned} \tag{7.30}$$

We have

$$\begin{aligned}
&\Big|\frac{N[d_{X_k}](x, r_j \pm \lambda^{\pm}_{j,k})}{(|X_k|-1)\widehat{\mu}_\rho[\rho](x, r_j)} - \widehat{\mathrm{vol}}[d_{X_k}, \rho](x, r_j \pm \lambda^{\pm}_{j,k})\Big| \\
&\quad\le \Big|\frac{1}{\widehat{\mu}_\rho[\rho](x, r_j)} - \frac{1}{\widehat{\mu}_\rho[\rho](x, r_j \pm \lambda^{\pm}_{j,k})}\Big| \\
&\quad\le \Big|\frac{1}{\widehat{\mu}_\rho[\rho](x, r_j)} - \frac{1}{\rho(x)}\Big| + \Big|\frac{1}{\widehat{\mu}_\rho[\rho](x, r_j \pm \lambda^{\pm}_{j,k})} - \frac{1}{\rho(x)}\Big| \\
&\quad\le \frac{A(r_j)}{\min(\rho)^2} + \frac{A(r_j + \lambda^{+}_{j,k})}{\min(\rho)^2} \\
&\quad\le \frac{2A(r_j + \lambda^{+}_{j,k})}{\min(\rho)^2}\,.
\end{aligned}$$

Therefore, by Eq. 7.30,

$$\begin{aligned}
\widehat{\mathrm{vol}}[d_{X_k}, \rho](x, r_j - \lambda^{-}_{j,k}) - \Big(\frac{\eta_k}{(\min(\rho)-\eta_k)^2} + \frac{2A(r_j)}{\min(\rho)^2} + \frac{2A(r_j + \lambda^{+}_{j,k})}{\min(\rho)^2}\Big) &\le \widehat{\mathrm{vol}}[\widehat{d_{X_k}}, \hat{\rho}](x, r_j) \\
&\le \widehat{\mathrm{vol}}[d_{X_k}, \rho](x, r_j + \lambda^{+}_{j,k}) + \Big(\frac{\eta_k}{(\min(\rho)-\eta_k)^2} + \frac{2A(r_j)}{\min(\rho)^2} + \frac{2A(r_j + \lambda^{+}_{j,k})}{\min(\rho)^2}\Big)\,.
\end{aligned}$$

Because $A(r)$ increases monotonically,

$$\begin{aligned}
\widehat{\mathrm{vol}}[d_{X_k}, \rho](x, r_j - \lambda^{-}_{j,k}) - \Big(\frac{\eta_k}{(\min(\rho)-\eta_k)^2} + \frac{4A(r_j + \lambda^{+}_{j,k})}{\min(\rho)^2}\Big) &\le \widehat{\mathrm{vol}}[\widehat{d_{X_k}}, \hat{\rho}](x, r_j) \\
&\le \widehat{\mathrm{vol}}[d_{X_k}, \rho](x, r_j + \lambda^{+}_{j,k}) + \Big(\frac{\eta_k}{(\min(\rho)-\eta_k)^2} + \frac{4A(r_j + \lambda^{+}_{j,k})}{\min(\rho)^2}\Big)\,. \qquad (7.31)
\end{aligned}$$

Next, we use Equation 7.31 to compare $\hat{y}_k(x, r_j)$ to $\hat{y}(x, r_j \pm \lambda^{\pm}_{j,k})$ for all $j \in \{1, \ldots, m_k\}$. Dividing Equation 7.31 by $v_n r_j^n$, we obtain

$$\begin{aligned}
\frac{\widehat{\mathrm{vol}}[d_{X_k}, \rho](x, r_j - \lambda^{-}_{j,k})}{v_n r_j^n} - \Big(\frac{\eta_k}{(\min(\rho)-\eta_k)^2 v_n r_j^n} + \frac{4A(r_j + \lambda^{+}_{j,k})}{v_n r_j^n \min(\rho)^2}\Big) &\le \hat{y}_k(x, r_j) \\
&\le \frac{\widehat{\mathrm{vol}}[d_{X_k}, \rho](x, r_j + \lambda^{+}_{j,k})}{v_n r_j^n} + \Big(\frac{\eta_k}{(\min(\rho)-\eta_k)^2 v_n r_j^n} + \frac{4A(r_j + \lambda^{+}_{j,k})}{v_n r_j^n \min(\rho)^2}\Big) \qquad (7.32)
\end{aligned}$$

for all j. We now compare $\widehat{\mathrm{vol}}[d_{X_k},\rho](x, r_j \pm \lambda_{j,k}^{\pm})/(v_n r_j^n)$ to $\hat{y}(x, r_j \pm \lambda_{j,k}^{\pm})$. We have

$$\begin{aligned}
\Big|\hat{y}(x, r_j \pm \lambda_{j,k}^{\pm}) - \frac{\widehat{\mathrm{vol}}[d_{X_k},\rho](x, r_j \pm \lambda_{j,k}^{\pm})}{v_n r_j^n}\Big| &= \frac{\widehat{\mathrm{vol}}[d_{X_k},\rho](x, r_j \pm \lambda_{j,k}^{\pm})}{v_n}\left|\frac{1}{(r_j \pm \lambda_{j,k}^{\pm})^n} - \frac{1}{r^n}\right| \\
&= \frac{N[d_{X_k}](x, r_j \pm \lambda_{j,k}^{\pm})}{(|X_k|-1)v_n \widehat{\mu}_\rho[\rho](x,r)}\left|\frac{1}{(r_j \pm \lambda_{j,k}^{\pm})^n} - \frac{1}{r_j^n}\right| \\
&\le \frac{1}{v_n \widehat{\mu}_\rho[\rho](x,r)}\Big|\frac{1}{(r_j \pm \lambda_{j,k}^{\pm})^n} - \frac{1}{r_j^n}\Big| \\
&\le \frac{1}{v_n \min(\rho)}\Big|\frac{1}{(r_j \pm \lambda_{j,k}^{\pm})^n} - \frac{1}{r_j^n}\Big| .
\end{aligned}$$

Because $g(r) = 1/r^n$ is convex and monotonically decreasing for $r > 0$, we have

$$\Big|\frac{1}{(r_j + \lambda_{j,k}^{+})^n} - \frac{1}{r_j^n}\Big| \le \lambda_{j,k}^{+}|g'(r_j)| = \frac{n\lambda_{j,k}^{+}}{r_j^{n+1}}$$

and

$$\Big|\frac{1}{(r_j - \lambda_{j,k}^{-})^n} - \frac{1}{r_j^n}\Big| \le \lambda_{j,k}^{-}|g'(r_j - \lambda_{j,k}^{-})| = \frac{n\lambda_{j,k}^{-}}{(r_j - \lambda_{j,k}^{-})^{n+1}} .$$

Therefore,

$$\Big|\hat{y}(x, r_j + \lambda_{j,k}^{+}) - \frac{\widehat{\mathrm{vol}}[d_{X_k},\rho](x, r_j + \lambda_{j,k}^{+})}{v_n r_j^n}\Big| \le \frac{1}{v_n \min(\rho)}\frac{n\lambda_{j,k}^{+}}{r_j^{n+1}} \tag{7.33}$$

and

$$\Big|\hat{y}(x, r_j - \lambda_{j,k}^{-}) - \frac{\widehat{\mathrm{vol}}[d_{X_k},\rho](x, r_j - \lambda_{j,k}^{-})}{v_n r_j^n}\Big| \le \frac{1}{v_n \min(\rho)}\frac{n\lambda_{j,k}^{-}}{(r_j - \lambda_{j,k}^{-})^{n+1}} . \tag{7.34}$$

Together, Equations 7.32, 7.33, and 7.34 show that

$$\begin{aligned}
\hat{y}_k(x, r_j) \ge \hat{y}(x, r_j - \lambda_{j,k}^{-}) - \Bigg(&\frac{\eta_k}{(\min(\rho) - \eta_k)^2 v_n r_j^n} + \frac{n\lambda_{j,k}^{-}}{v_n \min(\rho)(r_j - \lambda_{j,k}^{-})^{n+1}} \\
&+ \frac{4A(r_j + \lambda_{j,k}^{+})}{v_n r_j^n \min(\rho)^2}\Bigg), \\
\hat{y}_k(x, r_j) \le \hat{y}(x, r_j + \lambda_{j,k}^{+}) + \Bigg(&\frac{\eta_k}{(\min(\rho) - \eta_k)^2 v_n r_j^n} + \frac{n\lambda_{j,k}^{+}}{v_n \min(\rho) r_j^{n+1}} + \frac{4A(r_j + \lambda_{j,k}^{+})}{v_n r_j^n \min(\rho)^2}\Bigg).
\end{aligned}$$

We define the following error terms:

$$\xi_{j,k}^{+} := \frac{\eta_k}{(\min(\rho) - \eta_k)^2 v_n r_j^n} + \frac{n\lambda_{j,k}^{+}}{v_n \min(\rho) r_j^{n+1}} + \frac{4A(r_j + \lambda_{j,k}^{+})}{v_n r_j^n \min(\rho)^2},$$

$$\xi_{j,k}^{-} := \frac{\eta_k}{(\min(\rho) - \eta_k)^2 v_n r_j^n} + \frac{n\lambda_{j,k}^{-}}{v_n \min(\rho)(r_j - \lambda_{j,k}^{-})^{n+1}} + \frac{4A(r_j + \lambda_{j,k}^{+})}{v_n r_j^n \min(\rho)^2},$$

$$\xi_k := \max_j \{\xi_{j,k}^{+}, \xi_{j,k}^{-}\}.$$

The error terms $\xi_{j,k}^{\pm}$ are nonnegative. To complete the proof, it suffices to show that $\xi_k / r_{\max,\,k}^2 \to 0$ as $k \to \infty$. For sufficiently large k,

$$\frac{\eta_k}{v_n r_j^n (\min(\rho) - \eta_k)^2 r_{\max,\,k}^2} \leq \frac{\eta_k}{v_n (r_{\min,\,k} + (\Delta r)_k)^n (\frac{1}{2}\min(\rho))^2 r_{\max,\,k}^2}.$$

Rearranging the terms on the right-hand side, we obtain

$$\frac{\eta_k}{v_n (r_{\min,\,k} + (\Delta r)_k)^n (\frac{1}{2}\rho)^2 r_{\max,\,k}^2}$$
$$= \frac{4}{v_n \min(\rho)^2} \left(\frac{\eta_k}{(r_{\min,\,k} + (\Delta r)_k)^{n+2/3}} \right) \left(\frac{(r_{\min,\,k} + (\Delta r)_k)^{2/3}}{r_{\max,\,k}^2} \right).$$

By hypothesis, the quantity above approaches 0 as $k \to \infty$, so $\max_j \frac{\eta_k}{v_n r_j^n (\min(\rho) - \eta_k)^2 r_{\max,\,k}^2} \to 0$ as $k \to \infty$. Additionally,

$$\frac{n\lambda_{j,k}^{+}}{v_n \min(\rho) r_j^{n+1} r_{\max,\,k}^2} \leq \frac{n(\delta_k + (\Delta r)_k)}{v_n \min(\rho)(r_{\min,\,k} + (\Delta r)_k - \delta_k)^{n+1} r_{\max,\,k}^2},$$

$$\frac{n\lambda_{j,k}^{-}}{v_n \min(\rho)(r_j - \lambda_{j,k}^{-})^{n+1} r_{\max,\,k}^2} \leq \frac{n(\delta_k + (\Delta r)_k)}{v_n \min(\rho)(r_{\min,\,k} + (\Delta r)_k - \delta_k)^{n+1} r_{\max,\,k}^2}.$$

By hypothesis, the right-hand sides above approach 0 as $k \to \infty$. Finally, we upper bound $\frac{A(r_j + \lambda_{j,k}^{+})}{r_j^n r_{\max,\,k}^2}$ by recalling that $A(r)$ increases monotonically and

$$\begin{aligned} r_j + \lambda_{j,k}^{+} &= r_j + \ell_k (\Delta r)_k \\ &\leq r_j + \delta_k + (\Delta r)_k \\ &\leq 2r_j \end{aligned}$$

for sufficiently large k. Therefore, $\frac{A(r_j + \lambda_{j,k}^{+})}{r_j^n r_{\max,\,k}^2} \leq \frac{A(2r_j)}{r_{\max,\,k}^2 r_j^n}$, which approaches 0 by hypothesis. This implies that $\xi_k / r_{\max,\,k}^2 \to 0$ as $k \to \infty$. □

Lemma 7.A.5 (Stability of $\hat{C}$)**.** For each k, suppose that $\widehat{d_{X_k}}$ is a metric on X_k such that

$$\delta_k := \max_{x,x' \in X_k} |\widehat{d_{X_k}}(x,x') - d(x,x')| \to 0 \qquad \text{as } k \to \infty\,.$$

Suppose that $\hat{\rho}$ is a density estimator such that

$$\eta_k := \max_{x \in X_k} \left| \hat{\rho}[\widehat{d_{X_k}}](x) - \rho(x) \right| \to 0 \qquad \text{as } k \to \infty\,,$$

and suppose that $\hat{n}[\widehat{d_{X_k}}] = \hat{n}[d_{X_k}] = n$ for sufficiently large k. If the hyperparameter-value sequences satisfy

1. $\max_j \frac{A(2r_j)}{r_j^n r_{\max,\,k}^2} \to 0$ as $k \to \infty$,
2. $\eta_k / (r_{\min,\,k} + (\Delta r)_k)^{n+2/3} \to 0$ as $k \to \infty$,
3. $r_{\min,\,k} + (\Delta r)_k > \delta_k$ for sufficiently large k,
4. $|X_k|(\Delta r)_k (r_{\min,\,k} + (\Delta r)_k - \delta_k)^n \to \infty$ as $k \to \infty$,
5. $r_{\min,\,k} / r_{\max,\,k}^3 \to 0$ as $k \to \infty$,
6. $((\Delta r)_k + \delta_k)/r_{\max,\,k}^3 \to 0$ as $k \to \infty$, and
7. $((\Delta r)_k + \delta_k)/[(r_{\min,\,k} + (\Delta r)_k - \delta_k)^{n+1} r_{\max,\,k}^2] \to 0$ as $k \to \infty$

then $|\hat{C}[X_k, \widehat{d_{X_k}}, \hat{\rho}](x_k) - \hat{C}[X_k, d, \rho](x_k)| \to 0$ in probability as $k \to \infty$, where $\{x_k\}$ is any sequence of points such that $x_k \in X_k$.

Proof. To simplify our notation, we define

$$\begin{aligned}
\hat{C}_k(x) &:= \hat{C}[\widehat{d_{X_k}}, \hat{\rho}](x)\,, \\
\hat{C}(x) &:= \hat{C}[d, \rho](x)\,, \\
\hat{y}_k(x,r) &:= \hat{y}[\widehat{d_{X_k}}, \hat{\rho}](x,r)\,, \\
\hat{y}(x,r) &:= \hat{y}[d, \rho](x,r)
\end{aligned}$$

for all $x \in X_k$. Let $\ell_k = \min\{\ell'_k \in \mathbb{Z} \mid \ell'_k(\Delta r)_k \geq \delta_k\}$, let $a_k = r_{\min,\,k} + (\Delta r)_k - \delta_k$, and let $b_k = r_{\max,\,k} + \ell_k(\Delta r)_k$. By hypothesis and choice of ℓ_k,

$$\begin{aligned}
&a_k > 0 && \text{for all } k\,,\\
&|X_k|(a_k)^n \to \infty && \text{as } k \to \infty\,,\\
&a_k < (r_{\min,\,k} + (\Delta r)_k) \to 0 && \text{as } k \to \infty\,,\\
&b_k = r_{\max,\,k} + (\Delta r)_k + (\ell_k - 1)(\Delta r)_k < (r_{\max,\,k} + (\Delta r)_k + \delta_k) \to 0 && \text{as } k \to \infty\,.
\end{aligned}$$

Let $J := \{2 - \ell_k, \ldots, m_k + \ell_k\}$, where $m_k := \frac{r_{\max,\,k} - r_{\min,\,k}}{(\Delta r)_k}$ is the number of radial steps. Let $R_k := \{r_j \mid j \in J\} \cup \{r_{\min,\,k} + (\Delta r)_k + \delta_k\}$. We have

$$|R_k| = m_k + 2\ell_k \leq m_k + \frac{\delta_k}{(\Delta r)_k} + 1 \leq \frac{2}{(\Delta r)_k} + 1\,.$$

Because $|X_k|(\Delta r)_k(r_{\min,\,k} + (\Delta r)_k - \delta_k)^n \to \infty$, we have $\frac{|R_k|}{|X_k|a_k^n} \to 0$ as $k \to \infty$. Therefore, by Lemma 7.A.3,

$$\mathbb{P}\Big[\max_{r \in R_k} |\hat{y}(x, r) - 1| \leq 1\Big] \to 1 \tag{7.35}$$

as $k \to \infty$.

By Lemma 7.A.4, there is a nonnegative sequence $\{\xi_k\}$ such that $\xi_k / r_{\max,\,k}^2 \to 0$ and

$$\hat{y}[d_{X_k}, \rho](x_k, r_{j-\ell_k}) - \xi_k \leq \hat{y}[\widehat{d_{X_k}}, \hat{\rho}](x_k, r_j) \leq \hat{y}[d_{X_k}, \rho](x_k, r_{j+\ell_k}) + \xi_k \tag{7.36}$$

for all $j \geq 2$ and

$$\hat{y}[d_{X_k}, \rho](x_k, r_1 - \delta_k) - \xi_k \leq \hat{y}[\widehat{d_{X_k}}, \hat{\rho}](x_k, r_1) \leq \hat{y}[d_{X_k}, \rho](x_k, r_{1+\ell_k}) + \xi_k \tag{7.37}$$

for sufficiently large k. (The case $j = 1$ is different because it is not necessarily true that $r_{1-\ell_k} \geq 0$.)

Let $\epsilon > 0$. We want to show that $\mathbb{P}[|\hat{C}_k(x) - \hat{C}(x)| < \epsilon] \to 1$ as $k \to \infty$. By Equation 7.35, it suffices to show that $|\hat{C}_k(x) - \hat{C}(x)| < \epsilon$ for sufficiently large k if

$$\max_{r \in R_k} |\hat{y}(x, r) - 1| \leq 1\,. \tag{7.38}$$

Therefore, for the remainder of the proof, we assume Equation 7.38 holds.

First, we obtain an upper bound on $\hat{C}_k(x) - \hat{C}(x)$. The upper bounds in Equations 7.36–7.37 imply that

$$\begin{aligned}\hat{C}_k(x) &= \frac{5}{r_{\max,\,k}^5 - r_{\min,\,k}^5}\left(\sum_{j=1}^{m_k} r_j^2\Big(\hat{y}_k(x, r_j) - 1\Big)(\Delta r)_k\right) \\ &\leq \frac{5}{r_{\max,\,k}^5 - r_{\min,\,k}^5}\left(\sum_{j=1}^{m_k} r_j^2\Big(\hat{y}(x, r_{j+\ell_k}) - 1 + \xi_k\Big)(\Delta r)_k\right).\end{aligned}$$

Substituting $r_j^2 = r_{j+\ell_k}^2 - \ell_k(\Delta r)_k(2r_j + \ell_k(\Delta r)_k)$, we obtain

$$\begin{aligned}\hat{C}_k(x) &\leq \frac{5}{r_{\max,\,k}^5 - r_{\min,\,k}^5}\sum_{j=1}^{m_k}\Big(r_{j+\ell_k}^2 - \ell_k(\Delta r)_k(2r_j + \ell_k(\Delta r)_k)\Big)\Big(\hat{y}(x, r_{j+\ell_k}) - 1 + \xi_k\Big)(\Delta r)_k \\ &= \frac{5}{r_{\max,\,k}^5 - r_{\min,\,k}^5}\left(\sum_{j=1+\ell_k}^{m_k+\ell_k} r_j^2\Big(\hat{y}(x, r_j) - 1 + \xi_k\Big)(\Delta r)_k\right. \\ &\quad \left. - \sum_{j=1}^{m_k}\Big(\ell_k(\Delta r)_k(2r_j + \ell_k(\Delta r)_k)\Big)\Big(\hat{y}(x, r_{j+\ell_k}) - 1 + \xi_k\Big)(\Delta r)_k\right).\end{aligned}$$

Rearranging terms, we obtain

$$\begin{aligned}\hat{C}_k(x) \leq \hat{C}(x) + \frac{5}{r_{\max,\,k}^5 - r_{\min,\,k}^5}\Bigg(&\sum_{j=1+\ell_k}^{m_k+\ell_k} r_j^2\xi_k(\Delta r)_k + \sum_{j=m_k+1}^{m_k+\ell_k} r_j^2\Big(\hat{y}(x, r_j) - 1\Big)(\Delta r)_k \\ &- \sum_{j=1}^{\ell_k} r_j^2\Big(\hat{y}(x, r_j) - 1\Big)(\Delta r)_k \\ &- \sum_{j=1}^{m_k}\ell_k(\Delta r)_k(2r_j + \ell_k(\Delta r)_k)\Big(\hat{y}(x, r_{j+\ell_k}) - 1 + \xi_k\Big)(\Delta r)_k\Bigg).\end{aligned}$$

By Equation 7.38,

$$\begin{aligned}\hat{C}_k(x) \leq \hat{C}(x) &+ \frac{5\xi_k}{r_{\max,\,k}^5 - r_{\min,\,k}^5}\sum_{j=1+\ell_k}^{m_k+\ell_k} r_j^2(\Delta r)_k \\ &+ \frac{5(1+\xi_k)}{r_{\max,\,k}^5 - r_{\min,\,k}^5}\Bigg(\sum_{j=m_k+1}^{m_k+\ell_k} r_j^2(\Delta r)_k + \sum_{j=1}^{\ell_k} r_j^2(\Delta r)_k \\ &\qquad + \sum_{j=1}^{m_k}\ell_k(\Delta r)_k(2r_j + \ell_k(\Delta r)_k)(\Delta r)_k\Bigg).\end{aligned}$$

By comparing the sum $\sum_{j=1+\ell_k}^{m_k+\ell_k} r_j^2(\Delta r)_k$ to the integral $\int_0^{r_{\max,\,k}+(\Delta r)_k+\delta_k} r^2 dr$, we obtain

$$\hat{C}_k(x) \le \hat{C}(x) + \frac{5\xi_k}{r_{\max,\,k}^5 - r_{\min,\,k}^5} \cdot \frac{(r_{\max,\,k} + (\Delta r)_k + \delta_k)^3}{3} + \frac{5(1+\xi_k)}{r_{\max,\,k}^5 - r_{\min,\,k}^5} \left(\sum_{j=m_k+1}^{m_k+\ell_k} r_j^2(\Delta r)_k + \sum_{j=1}^{\ell_k} r_j^2(\Delta r)_k + \sum_{j=1}^{m_k} \ell_k(\Delta r)_k(2r_j + \ell_k(\Delta r)_k)(\Delta r)_k \right).$$

By hypothesis, $r_{\min,\,k} \le B r_{\max,\,k}$ for some $B < 1$ and $(\Delta r)_k + \delta_k < r_{\max,\,k}$ for sufficiently large k, so

$$\hat{C}_k(x) \le \hat{C}(x) + \frac{40}{3(1-B^5)} \cdot \frac{\xi_k}{r_{\max,\,k}^2} + \frac{5(1+\xi_k)}{(1-B^5) r_{\max,\,k}^5} \left(\sum_{j=m_k+1}^{m_k+\ell_k} r_j^2(\Delta r)_k + \sum_{j=1}^{\ell_k} r_j^2(\Delta r)_k + \sum_{j=1}^{m_k} \ell_k(\Delta r)_k(2r_j + \ell_k(\Delta r)_k)(\Delta r)_k \right).$$

Because r_j increases monotonically with j,

$$\hat{C}_k(x) \le \hat{C}(x) + \frac{40}{3(1-B^5)} \cdot \frac{\xi_k}{r_{\max,\,k}^2} + \frac{5(1+\xi_k)}{(1-B^5) r_{\max,\,k}^5} \Big(\ell_k(\Delta r)_k (r_{\max,\,k} + \ell_k(\Delta r)_k)^2 + \ell_k(\Delta r)_k (r_{\min,\,k} + \ell_k(\Delta r)_k)^2 + (r_{\max,\,k} - r_{\min})\ell_k(\Delta r)_k(2r_{\max,\,k} + \ell_k(\Delta r)_k) \Big).$$

By choice of ℓ_k, we have $\ell_k(\Delta r)_k < \delta_k + (\Delta r)_k$, so

$$\hat{C}_k(x) \le \hat{C}(x) + \frac{40}{3(1-B^5)} \cdot \frac{\xi_k}{r_{\max,\,k}^2} + \frac{5(1+\xi_k)(\delta_k + (\Delta r)_k)}{(1-B^5) r_{\max,\,k}^5} \Big((r_{\max,\,k} + (\delta_k + (\Delta r)_k))^2 + (r_{\min,\,k} + (\delta_k + (\Delta r)_k))^2 + (r_{\max,\,k} - r_{\min,\,k})(2r_{\max,\,k} + (\delta_k + (\Delta r)_k)) \Big).$$

Because $0 \le r_{\min,\,k} < r_{\max,\,k}$,

$$\hat{C}_k(x) \le \hat{C}(x) + \frac{40}{3(1-B^5)} \cdot \frac{\xi_k}{r_{\max,\,k}^2} + \frac{5(1+\xi_k)(\delta_k + (\Delta r)_k)}{(1-B^5) r_{\max,\,k}^5} \Big(2(r_{\max,\,k} + (\delta_k + (\Delta r)_k))^2 + r_{\max,\,k}(2r_{\max,\,k} + \delta_k + (\Delta r)_k) \Big).$$

By hypothesis, $\delta_k + (\Delta r)_k < r_{\max,\,k}$ for sufficiently large k, so

$$\hat{C}_k(x) - \hat{C}(x) \leq \frac{40}{3(1-B^5)} \cdot \frac{\xi_k}{r_{\max,\,k}^2} + \frac{55(1+\xi_k)}{(1-B^5)} \cdot \frac{(\delta_k + (\Delta r)_k)}{r_{\max,\,k}^3}$$

for sufficiently large k. The right-hand side is positive and approaches 0 as $k \to \infty$, so

$$\hat{C}_k(x) - \hat{C}(x) < \epsilon \tag{7.39}$$

for sufficiently large k.

Next, we obtain a lower bound on $\hat{C}_k(x) - \hat{C}(x)$. The calculation proceeds almost the same way as our calculation of an upper bound, except that the lower bound in Equation 7.37 is of a slightly different form than the upper bound. The lower bounds in Equations 7.36–7.37 imply that

$$\begin{aligned}
\hat{C}_k(x) &= \frac{5}{r_{\max,\,k}^5 - r_{\min,\,k}^5} \left(\sum_{j=1}^{m_k} r_j^2 \Big(\hat{y}_k(x, r_j) - 1 \Big) (\Delta r)_k \right) \\
&\geq \frac{5}{r_{\max,\,k}^5 - r_{\min,\,k}^5} \left(r_1^2 \Big(\hat{y}_k(x, r_1 - \delta_k) - 1 - \xi_k \Big) (\Delta r)_k \right. \\
&\qquad \left. + \sum_{j=2}^{m_k} r_j^2 \Big(\hat{y}(x, r_{j-\ell_k}) - 1 - \xi_k \Big) (\Delta r)_k \right).
\end{aligned}$$

Substituting $r_j^2 = r_{j-\ell_k}^2 + \ell_k (\Delta r)_k (2r_j - \ell_k (\Delta r)_k)$, we obtain

$$\begin{aligned}
\hat{C}_k(x) \geq \frac{5}{r_{\max,\,k}^5 - r_{\min,\,k}^5} \Bigg(& r_1^2 \Big(\hat{y}(x, r_1 - \delta_k) - 1 - \xi_k \Big) (\Delta r)_k \\
& + \sum_{j=2-\ell_k}^{m_k - \ell_k} r_j^2 \Big(\hat{y}(x, r_j) - 1 - \xi_k \Big) (\Delta r)_k \\
& + \sum_{j=2}^{m_k} \ell_k (\Delta r)_k (2r_j - \ell_k (\Delta r)_k) \Big(\hat{y}(x, r_{j-\ell_k}) - 1 - \xi_k \Big) (\Delta r)_k \Bigg).
\end{aligned}$$

Rearranging terms, we have

$$\hat{C}_k(x) \geq \hat{C}(x) + \frac{5}{r_{\max,k}^5 - r_{\min,k}^5}\Bigg(r_1^2\Big(\hat{y}(x, r_1 - \delta_k) - 1 - \xi_k\Big)(\Delta r)_k + \sum_{j=2-\ell_k}^{0} r_j^2\Big(\hat{y}(x, r_j) - 1\Big)(\Delta r)_k - r_1^2\Big(\hat{y}(x, r_1) - 1\Big)(\Delta r)_k - \sum_{j=m_k-\ell_k+1}^{m_k} r_j^2\Big(\hat{y}(x, r_j) - 1\Big)(\Delta r)_k - \sum_{j=2-\ell_k}^{m_k-\ell_k} r_j^2 \xi_k (\Delta r)_k + \sum_{j=2}^{m_k} \ell_k (\Delta r)_k (2r_j - \ell_k(\Delta r)_k)\Big(\hat{y}(x, r_{j-\ell_k}) - 1 - \xi_k\Big)(\Delta r)_k\Bigg).$$

By Equation 7.38,

$$\hat{C}_k(x) \geq \hat{C}(x) - \frac{5\xi_k}{r_{\max,k}^5 - r_{\min,k}^5} \sum_{j=2-\ell_k}^{m_k-\ell_k} r_j^2 (\Delta r)_k - \frac{5(1+\xi_k)}{r_{\max,k}^5 - r_{\min,k}^5}\Bigg(r_1^2(\Delta r)_k + \sum_{j=2-\ell_k}^{1} r_j^2(\Delta r)_k + \sum_{j=m_k-\ell_k+1}^{m_k} r_j^2(\Delta r)_k + \sum_{j=2}^{m_k} \ell_k(\Delta r)_k(2r_j - \ell_k(\Delta r)_k)(\Delta r)_k\Bigg).$$

By comparing the sum $\sum_{j=2-\ell_k}^{m_k-\ell_k} r_j^2(\Delta r)_k$ to the integral $\int_0^{r_{\max}} r^2 dr$, we obtain

$$\hat{C}_k(x) \geq \hat{C}(x) - \left(\frac{5\xi_k}{r_{\max,k}^5 - r_{\min,k}^5} \cdot \frac{r_{\max,k}^3}{3}\right) - \frac{5(1+\xi_k)}{r_{\max,k}^5 - r_{\min,k}^5}\Bigg(r_1^2(\Delta r)_k + \sum_{j=2-\ell_k}^{1} r_j^2(\Delta r)_k + \sum_{j=m_k-\ell_k+1}^{m_k} r_j^2(\Delta r)_k + \sum_{j=2}^{m_k} \ell_k(\Delta r)_k(2r_j - \ell_k(\Delta r)_k)(\Delta r)_k\Bigg).$$

By hypothesis, $r_{\min,k} \leq B r_{\max,k}$ for some $B < 1$, so

$$\hat{C}_k(x) \geq \hat{C}(x) - \left(\frac{5}{3(1-B^5)} \cdot \frac{\xi_k}{r_{\max,k}^2}\right) - \frac{5(1+\xi_k)}{(1-B^5) r_{\max,k}^5}\Bigg(r_1^2(\Delta r)_k + \sum_{j=2-\ell_k}^{1} r_j^2(\Delta r)_k + \sum_{j=m_k-\ell_k+1}^{m_k} r_j^2(\Delta r)_k + \sum_{j=2}^{m_k} \ell_k(\Delta r)_k(2r_j - \ell_k(\Delta r)_k)(\Delta r)_k\Bigg).$$

Because r_j^2 increases monotonically with $j \in \{2-\ell_k, \ldots, m_k\}$ (and noting that $r_{2-\ell_k} > 0$ by

hypothesis),

$$\begin{aligned}
\hat{C}_k(x) &\geq \hat{C}(x) - \left(\frac{5}{3(1-B^5)} \cdot \frac{\xi_k}{r_{\max,\,k}^2}\right) - \frac{5(1+\xi_k)}{(1-B^5)r_{\max,\,k}^5}\Bigg(r_1^2(\Delta r)_k + \ell_k r_1^2(\Delta r)_k \\
&\qquad + \ell_k r_{\max,\,k}^2(\Delta r)_k + (m_k-1)(\Delta r)_k\ell_k(\Delta r)_k(2r_{\max,\,k} - \ell_k(\Delta r)_k)\Bigg). \\
&\geq \hat{C}(x) - \frac{5}{3(1-B^5)} \cdot \frac{\xi_k}{r_{\max,\,k}^2} - \frac{5(1+\xi_k)}{(1-B^5)r_{\max,\,k}^5}\Big((1+\ell_k)r_1^2(\Delta r)_k + 3r_{\max,\,k}^2(\Delta r)_k\ell_k\Big) \\
&\geq \hat{C}(x) - \frac{5}{3(1-B^5)} \cdot \frac{\xi_k}{r_{\max,\,k}^2} \\
&\qquad - \frac{5(1+\xi_k)}{(1-B^5)r_{\max,\,k}^5}\Big((1+2\ell_k)r_{\max,\,k}^2(\Delta r)_k + 3r_{\max,\,k}^2(\Delta r)_k\ell_k\Big).
\end{aligned}$$

By choice of ℓ_k, we have $\ell_k(\Delta r)_k < (\Delta r)_k + \delta_k$, which implies

$$\begin{aligned}
\hat{C}_k(x) - \hat{C}(x) &\geq -\left(\frac{5}{3(1-B^5)} \cdot \frac{\xi_k}{r_{\max,\,k}^2}\right) - \frac{5(1+\xi_k)}{(1-B^5)r_{\max,\,k}^3}\Big(6(\Delta r)_k + 5\delta_k\Big) \\
&\geq -\left(\frac{5}{3(1-B^5)} \cdot \frac{\xi_k}{r_{\max,\,k}^2}\right) - \frac{30(1+\xi_k)}{(1-B^5)} \cdot \frac{(\Delta r)_k + \delta_k}{r_{\max,\,k}^3}.
\end{aligned}$$

The right-hand side is negative and approaches 0 as $k \to \infty$, so

$$\hat{C}_k(x) - \hat{C}(x) > -\epsilon \tag{7.40}$$

for sufficiently large k. Together, Equations 7.39 and 7.40 complete the proof. □

Lemma 7.A.6. If the hyperparameter-value sequences satisfy

1. $(\Delta r)_k / r_{\max,\,k}^3 \to 0$ as $k \to \infty$,
2. $|X_k|(r_{\min,\,k} + (\Delta r)_k)^n \to \infty$, and
3. $r_{\min,\,k}/r_{\max,\,k}^3 \to 0$ as $k \to \infty$,

then $|\hat{C}[d_{X_k}, \rho](x_k) - C(x_k)| \to 0$ in probability as $k \to \infty$, where $\{x_k\}$ is any sequence of points such that $x_k \in X_k$.

Proof. Let x be any point in X_k. For all $j \in \{1, \ldots, m_k\}$, let $\hat{y}_j := \hat{y}(x, r_j)$ and let $y_j := y(x, r_j)$. The absolute difference $|\hat{C}(x) - C(x)|$ is bounded above by

$$|\hat{C}(x) - C(x)| \leq \frac{5}{r_{\max,\,k}^5 - r_{\min,\,k}^5} \left(\Big| \sum_{j=1}^{m_k} r_j^2 (\Delta r)_k - \int_{r_{\min,\,k}}^{r_{\max,\,k}} r^2 dr \Big| + \Big| \int_{r_{\min,\,k}}^{r_{\max,\,k}} r^2 y(x,r) dr - \sum_{j=1}^{m_k} r_j^2 y_j (\Delta r)_k \Big| + \Big| \sum_{j=1}^{m_k} r_j^2 (\hat{y}_j - y_j)(\Delta r)_k \Big| \right).$$

Because $r_{\min,\,k}/r_{\max,\,k}^3 \to 0$ (by hypothesis), there is a constant $B < 1$ such that $r_{\min,\,k} \leq B r_{\max,\,k}$ for all k. Therefore,

$$|\hat{C}(x) - C(x)| \leq \frac{5}{(1 - B^5) r_{\max,\,k}^5} \left(\Big| \sum_{j=1}^{m_k} r_j^2 (\Delta r)_k - \int_{r_{\min,\,k}}^{r_{\max,\,k}} r^2 dr \Big| + \Big| \int_{r_{\min,\,k}}^{r_{\max,\,k}} r^2 y(x,r) dr - \sum_{j=1}^{m_k} r_j^2 y_j (\Delta r)_k \Big| + \Big| \sum_{j=1}^{m_k} r_j^2 (\hat{y}_j - y_j)(\Delta r)_k \Big| \right). \tag{7.41}$$

The first term on the right-hand side of Equation 7.41 is a Riemann-sum error. For any function $f(r)$ that is integrated on $[r_{\min,\,k}, r_{\max,\,k}]$, the error in the right Riemann sum is bounded above by $\max_{r \in [r_{\min,\,k}, r_{\max,\,k}]} |f'(r)| (\Delta r)_k \cdot (r_{\max,\,k} - r_{\min,\,k})/2$. Therefore,

$$\begin{aligned} &\Big| \int_{r_{\min,\,k}}^{r_{\max,\,k}} r^2 y(x,r) dr - \sum_{j=1}^{m_k} r_j^2 y_j (\Delta r)_k \Big| \\ &\quad \leq (\Delta r)_k \Big(\max_{r \in [r_{\min,\,k}, r_{\max,\,k}]} \Big| \frac{d}{dr} r^2 y(x,r) \Big| \Big) (r_{\max,\,k} - r_{\min,\,k})/2 \\ &\quad \leq (\Delta r)_k \Big(\max_{r \in [r_{\min,\,k}, r_{\max,\,k}]} \Big| \frac{d}{dr} r^2 y(x,r) \Big| \Big) r_{\max,\,k}/2. \end{aligned} \tag{7.42}$$

We have $\frac{d}{dr} r^2 y(x,r) = r^2 \frac{d}{dr} y(x,r) + 2ry(x,r)$. By Equation 7.1, we have $\lim_{r \to 0} y(x,r) = 1$ and $\lim_{r \to 0} \frac{d}{dr} y(x,r) = 0$. Therefore, $|\frac{d}{dr} r^2 y(x,r)| = \mathcal{O}(r)$ as $r \to 0$, so there is a constant $A > 1$ such that $\max_{r \in [r_{\min,\,k}, r_{\max,\,k}]} |\frac{d}{dr} r^2 y(x,r)| \leq 2A r_{\max,\,k}$ for sufficiently small $r_{\max,\,k}$. Thus for sufficiently large k,

$$\Big(\max_{r \in [r_{\min,\,k}, r_{\max,\,k}]} \Big| \frac{d}{dr} r^2 y(x,r) \Big| \Big) r_{\max,\,k}/2 \leq A r_{\max,\,k}^2 \tag{7.43}$$

because $r_{\max,\,k} \to 0$ as $k \to \infty$.

Let $\epsilon > 0$. By hypothesis, $\frac{r^3_{\max,\,k}}{(\Delta r)_k} \to \infty$ as $k \to \infty$, so

$$r^2_{\max,\,k} \leq \frac{r^5_{\max,\,k}(1 - B^5)\epsilon}{15A(\Delta r)_k} \tag{7.44}$$

for sufficiently large k. Substituting Equation 7.44 into Equation 7.43 and Eq. 7.43 into Eq. 7.42 yields

$$\Big| \int_{r_{\min,\,k}}^{r_{\max,\,k}} r^2 y(x, r) dr - \sum_{j=1}^{m_k} r_j^2 y_j (\Delta r)_k \Big| \leq \frac{r^5_{\max,\,k}(1 - B^5)\epsilon}{15}\,. \tag{7.45}$$

Next, we bound the second term on the right-hand side of Equation 7.41, which is also a Riemann-sum error. For a monotonic function $f(r)$ that is integrated on $[r_{\min,\,k}, r_{\max,\,k}]$, the error in the right Riemann sum is bounded above by $(\Delta r)_k |f(r_{\max,\,k}) - f(r_{\min,\,k})|$. Therefore,

$$\Big| \sum_{j=1}^{m_k} r_j^2 (\Delta r)_k - \int_{r_{\min,\,k}}^{r_{\max,\,k}} r^2 dr \Big| \leq (\Delta r)_k (r^2_{\max,\,k} - r^2_{\min,\,k}) \leq (\Delta r)_k \cdot r^2_{\max,\,k}\,.$$

By Eq. 7.44,

$$\Big| \sum_{j=1}^{m_k} r_j^2 (\Delta r)_k - \int_{r_{\min,\,k}}^{r_{\max,\,k}} r^2 dr \Big| \leq \frac{r^5_{\max,\,k}(1 - B^5)\epsilon}{15A} < \frac{r^5_{\max,\,k}(1 - B^5)\epsilon}{15} \tag{7.46}$$

for sufficiently large k.

Putting the inequalities of Eqs. 7.46 and 7.45 into Eq. 7.41, we obtain

$$|\hat{C}(x) - C(x)| \leq \frac{2}{3}\epsilon + \Big| \sum_{j=1}^{m_k} r_j^2 (\hat{y}_j - y_j) \Big| \frac{5(\Delta r)_k}{(1 - B^5) r^5_{\max,\,k}}.$$

Therefore,

$$\mathbb{P}[|\hat{C}(x) - C(x)| > \epsilon] \leq \mathbb{P}\left[\Big| \sum_{j=1}^{m_k} r_j^2 (\hat{y}_j - y_j) \Big| > \frac{(1 - B^5) r^5_{\max,\,k} \epsilon}{15(\Delta r)_k} \right]. \tag{7.47}$$

We have $\mathbb{E}\Big[\sum_{j=1}^{m_k} r_j^2 \hat{y}_j \Big] = \sum_{j=1}^{m_k} r_j^2 y_j$ because $\mathbb{E}[\hat{y}_j] = y_j$ (Lemma 7.2.4). By applying Chebyshev's inequality to the right-hand side of Eq. 7.47, we obtain

$$\mathbb{P}[|\hat{C}(x) - C(x)| > \epsilon] \leq \Big(\frac{15}{(1 - B^5)\epsilon} \Big)^2 \Big(\frac{(\Delta r)_k}{r^5_{\max,\,k}} \Big)^2 \mathrm{var}\Big(\sum_{j=1}^{m_k} r_j^2 \hat{y}_j \Big). \tag{7.48}$$

We expand the variance as

$$\mathrm{var}\Big(\sum_{j=1}^{m_k} r_j^2 \hat{y}_j\Big) = \sum_{j=1}^{m_k} r_j^4 \mathrm{var}(\hat{y}_j) + \sum_{i \neq j} r_i^2 r_j^2 \mathrm{cov}(\hat{y}_i, \hat{y}_j).$$

For all $i \neq j$, we have $\mathrm{cov}(\hat{y}_i, \hat{y}_j)^2 \leq \mathrm{var}(\hat{y}_i)\mathrm{var}(\hat{y}_j)$. Therefore,

$$\mathrm{var}\Big(\sum_{j=1}^{m_k} r_j^2 \hat{y}_j\Big) \leq \Big(\sum_{j=1}^{m_k} r_j^2 \sqrt{\mathrm{var}(\hat{y}_j)}\Big)^2.$$

By Prop. 7.4.1, there is a constant $A' \geq 0$ such that

$$\mathrm{var}(\hat{y}_j) \leq \frac{A'}{|X_k| r_j^n}$$

for all j and sufficiently large k. Therefore,

$$\mathrm{var}\Big(\sum_{j=1}^{m_k} r_j^2 \hat{y}_j\Big) \leq \frac{A'}{|X_k|}\Big(\sum_{j=1}^{m_k} r_j^{2-n/2}\Big)^2 \tag{7.49}$$

for sufficiently large k. Below, we use Eq. 7.49 to obtain an upper bound on the right-hand side of Eq. 7.48. There are two cases, depending on n.

Case 1: $(2 \leq n \leq 4)$.

In this case,

$$\Big(\sum_{j=1}^{m_k} r_j^{2-n/2}\Big)^2 \leq r_{\max,\,k}^{4-n}\Big(\frac{r_{\max,\,k} - r_{\min,\,k}}{(\Delta r)_k}\Big)^2 \leq \frac{r_{\max,\,k}^{6-n}}{(\Delta r)_k^2} \tag{7.50}$$

because $r_j^{2-n/2}$ is monotonically increasing. Combining Eqs 7.48, 7.49, and 7.50, we obtain

$$\begin{aligned}
\mathbb{P}[|\hat{C}(x) - C(x)| > \epsilon] &\leq \Big(\frac{15}{(1-B^5)\epsilon}\Big)^2 \Big(\frac{(\Delta r)_k}{r_{\max,\,k}^5}\Big)^2 \frac{A'}{|X_k|} \frac{r_{\max,\,k}^{6-n}}{(\Delta r)_k^2} \\
&\leq A'\Big(\frac{15}{(1-B^5)\epsilon}\Big)^2 \frac{1}{|X_k| r_{\max,\,k}^{n+4}} \\
&= A'\Big(\frac{15}{(1-B^5)\epsilon}\Big)^2 \frac{1}{|X_k|(r_{\min,\,k} + (\Delta r)_k)^{n/3+4/3}} \Big(\frac{r_{\min,\,k} + (\Delta r)_k}{r_{\max,\,k}^3}\Big)^{n/3+4/3}.
\end{aligned}$$

Because $n/3 + 4/3 \leq n$ for $n \geq 2$ and $r_{\min,\,k} + (\Delta r)_k < 1$ for sufficiently large k,

$$\mathbb{P}[|\hat{C}(x) - C(x)| > \epsilon] \leq A'\Big(\frac{15}{(1-B^5)\epsilon}\Big)^2 \frac{1}{|X_k|(r_{\min,\,k} + (\Delta r)_k)^n} \Big(\frac{r_{\min,\,k} + (\Delta r)_k}{r_{\max,\,k}^3}\Big)^{n/3+3/4}$$

for sufficiently large k. By hypothesis, $\left(\frac{r_{\min,\,k}+(\Delta r)_k}{r^3_{\max,\,k}}\right) \to 0$ and $\frac{1}{|X_k|(r_{\min,\,k}+(\Delta r)_k)^n} \to 0$ as $k \to \infty$. Therefore, $\mathbb{P}[|\hat{C}(x_k) - C(x_k)| > \epsilon] \to 0$ as $k \to \infty$.

Case 2: $(n > 4)$.

In this case,

$$\Big(\sum_{j=1}^{m_k} r_j^{2-n/2}\Big)^2 \leq (r_{\min,\,k} + (\Delta r)_k)^{4-n}\Big(\frac{r_{\max,\,k} - r_{\min,\,k}}{(\Delta r)_k}\Big)^2 \leq (r_{\min,\,k} + (\Delta r)_k)^{4-n}\Big(\frac{r_{\max,\,k}}{(\Delta r)_k}\Big)^2 \tag{7.51}$$

because $r_j^{2-n/2}$ is monotonically decreasing. Combining Eqs 7.48, 7.49, and 7.51 yields

$$\begin{aligned}\mathbb{P}[|\hat{C}(x) - C(x)| > \epsilon] &\leq A'\Big(\frac{15}{(1-B^5)\epsilon}\Big)^2\Big(\frac{(\Delta r)_k}{r^5_{\max,\,k}}\Big)^2\frac{(r_{\min,\,k} + (\Delta r)_k)^{4-n} r^2_{\max,\,k}}{|X_k|(\Delta r)_k^2} \\ &= A'\Big(\frac{15}{(1-B^5)\epsilon}\Big)^2\Big(\frac{r_{\min,\,k} + (\Delta r)_k}{r^2_{\max,\,k}}\Big)^4\frac{1}{|X_k|(r_{\min,\,k} + (\Delta r)_k)^n}\,.\end{aligned}$$

By hypothesis, $\left(\frac{r_{\min,\,k}+(\Delta r)_k}{r^2_{\max,\,k}}\right) \to 0$ and $\frac{1}{|X_k|(r_{\min,\,k}+(\Delta r)_k)^n} \to 0$ as $k \to \infty$. Therefore, $\mathbb{P}[|\hat{C}(x_k) - C(x_k)| > \epsilon] \to 0$ as $k \to \infty$. □

Acknowledgements

We thank Yining Liu, Michael Mandell, and Mason Porter for helpful conversations.

CHAPTER 8

Conclusions

In this thesis, I presented new applications, methods, and theory of topological data analysis (TDA) and geometric data analysis (GDA).

First, I introduced the concept of a PD bundle, which is a generalization of a vineyard (which is used, e.g., for time-varying persistent homology) that allows one to analyze the topological features of any parameterized set of filtrations. I proved that generic PD bundles are determined by finitely many points in the "base space" (the parameter space) when the base space is a smooth compact manifold. Exploiting this fact, I developed an algorithm to compute "piecewise-linear" PD bundles. We also explored the phenomenom of "monodromy" that can occur in PD bundles but does not occur in vineyards. The development of PD bundles raises many questions. In particular, it remains to understand the stability of PD bundles, to develop good summary statistics for applications, and to strengthen the connection with important special cases like the persistent homology transform.

In the next part of the thesis, we applied TDA to several spatial and spatiotemporal data sets. The first set of applications (see Chapter 5) was a set of COVID-19 data sets (case-count data in Los Angeles and vaccination data in New York City), for which we used TDA to analyze both the local and global structure of local extrema. For the Los Angeles data, we used vineyards to study how the spatial structure of case-count anomalies (i.e., local maxima) evolved with time. In the next chapter (see Chapter 6), we used PH to study the accessibility of polling places and to identify "holes in coverage." In ongoing work, my collaborators and I are generalizing our method to study the accessibility of public green space (specifically,

parks) in urban areas. This extension requires mathematical modifications to account for the different quality of different parks.

In the last part of the thesis, we defined an intrinsic estimator for scalar curvature. The estimator can be applied not only to point-cloud data but also to any finite metric space, such as a network that is equipped with the shortest-path metric. We proved that the estimator converges to the ground-truth scalar curvature and that it is stable with respect to perturbations in the metric structure. We also validated it on synthetic data that is sampled from manifolds with known curvature. One of the broad challenges in geometric data analysis is that local geometric invariants (such as scalar curvature) are sensitive to noise and small perturbations of the underlying manifold. In our curvature-estimation research, we addressed this challenge by estimating the quantity of interest (geodesic-ball volume) at multiple length scales. This strategy is also the key behind the stability of PH. A major goal of future work is to develop strategies to stably estimate other local geometric invariants.

To conclude, this thesis strengthens the theoretical and computational foundations of TDA and GDA, which are fields that use ideas from algebraic topology, differential geometry, computational geometry, and statistics to analyze (typically) high-dimensional data. By harnessing the power of algebraic topology and differential geometry, we developed novel methods to reveal structure in complex data sets.

APPENDIX A

Opinion Dynamics

In [HKF22], I led a study of opinion dynamics on hypergraphs. The content of this paper is not included in the dissertation only because the topic does not fit with the other papers. In this appendix, I briefly summarize our results and contributions, adapted from [HKF22].

In opinion dynamics, one models how people's opinions evolve with time as they interact with each other. Typically, one models social relationships as edges of a network and nodes can interact only with their neighbors in the network. A bounded-confidence model (BCM) is a model of opinion dynamics in which nodes are influenced only by nodes whose opinions are sufficiently close (within a "confidence bound" c) of their own opinion. The opinion of each node is represented as a real number, which is an appropriate choice for opinions that lie on a one-dimensional continuous spectrum.

In [HKF22], we developed a BCM of opinion dynamics on hypergraphs, which allowed us to study the effect of polyadic social interactions on opinion formation. (Each hyperedge of a hypergraph encodes a polyadic social relationship.) We proved that our hypergraph BCM almost surely converges to consensus on the complete hypergraph for a wide range of initial conditions for the opinions of the nodes, including all bounded initial opinion distributions. We also showed that, under suitable conditions, echo chambers can form on hypergraphs with community structure. We demonstrated that the opinions of nodes can sometimes jump from one opinion cluster to another in a single time step; this phenomenon (which we called "opinion jumping") is not possible in BCMs on ordinary graphs. Additionally, we analyzed the convergence time of our model.

REFERENCES

[ACG19] Hirokazu Anai, Frédéric Chazal, Marc Glisse, Yuichi Ike, Hiroya Inakoshi, Raphaël Tinarrage, and Yuhei Umeda. "DTM-Based Filtrations." In Gill Barequet and Yusu Wang, editors, *35th International Symposium on Computational Geometry (SoCG 2019)*, volume 129 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 58:1–58:15, Dagstuhl, Germany, 2019. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[ACM22] Shreya Arya, Justin Curry, and Sayan Mukherjee. "A Sheaf-Theoretic Construction of Shape Space." *arXiv:2204.09020*, 2022.

[AEH06] Pankaj K. Agarwal, Herbert Edelsbrunner, John Harer, and Yusu Wang. "Extreme elevation on a 2-manifold." *Discrete & Computational Geometry*, **36**:553–572, 2006.

[AEK17] Henry Adams, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier. "Persistence Images: A Stable Vector Representation of Persistent Homology." *Journal of Machine Learning Research*, **18**(8):1–35, 2017.

[AHH19] G. Arvanitidis, S. Hauberg, P. Hennig, and M. Schober. "Fast and Robust Shortest Paths on Manifolds Learned from Data." In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89, pp. 1506–1515. PMLR, 2019.

[And13] Massimo Ferri Patrizio Frosini Claudia Landi Andrea Cerri, Barbara Di Fabio. "Betti Numbers in Multidimensional persistent homology are stable functions." *Mathematical Methods in the Applied Sciences*, **36**(12):1543–1557, 2013.

[AQO22] Erik J. Amézquita, Michelle Y. Quigley, Tim Ophelders, Jacob B. Landis, Daniel Koenig, Elizabeth Munch, and Daniel H. Chitwood. "Measuring hidden phenotype: quantifying the shape of barley seeds using the Euler characteristic transform." *in silico Plants*, **4**(1), 2022.

[Ari21] Julien Arino. "Describing, modelling and forecasting the spatial and temporal spread of COVID-19 — A short review." *Mathematics of Public Health*, pp. 25–51, 2021.

[BB04] Lars Brabyn and Ross Barnett. "Population need and geographical access to general practitioners in rural New Zealand." *New Zealand Med. J.*, **117**(1199):U996, 2004.

[BB06] Lars Brabyn and Paul Beere. "Population access to hospital emergency departments and the impacts of health reform in New Zealand." *Health Inform. J.*, **12**(3):227–237, 2006.

[BBG18] Hugo Barbosa, Marc Barthelemy, Gourab Ghoshal, Charlotte R. James, Maxime Lenormand, Thomas Louail, Ronaldo Menezes, José J. Ramasco, Filippo Simini, and Marcello Tomasini. "Human mobility: Models and applications." *Physics Reports*, **734**:1–74, 2018.

[BBH06] Raymond C. Browning, Emily A. Baker, Jessica A. Herron, and Rodger Kram. "Effects of obesity and sex on the energetic cost and preferred speed of walking." *J. Appl. Physiol.*, **100**(2):390–398, 2006.

[BCK08] Mark de Berg, Otfried Cheong, Marc van Kreveld, and Mark Overmars. *Computational Geometry: Algorithms and Applications*. Springer, Berlin, Heidelberg, 3rd edition, 2008.

[BH13] Dirk Brockmann and Dirk Helbing. "The Hidden Geometry of Complex, Network-Driven Contagion Phenomena." *Science*, **342**(6164):1337–1342, 2013.

[BHO18] Mickaël Buchet, Yasuaki Hiraoka, and Ippei Obayashi. "Persistent Homology and Materials Informatics." In Isao Tanaka, editor, *Nanoinformatics*, pp. 75–95. Springer, Singapore, 2018.

[BHP20] Peter Bubenik, Michael Hull, Dhruv Patel, and Benjamin Whittle. "Persistent homology detects curvature." *Inverse Problems*, **36**(2):025008, 2020.

[BL22] Magnus Bakke Botnan and Michael Lesnick. "An introduction to multiparameter persistence." *arxiv:2203.14289*, 2022.

[BLM19] Gregory Bell, Austin Lawson, Joshua Martin, James Rudzinski, and Clifford Smyth. "Weighted persistent homology." *Involve: A Journal of Mathematics*, **12**(5):823–837, 2019.

[BMF22] Dhananjay Bhaskar, Kincaid MacDonald, Oluwadamilola Fasina, Dawson S. Thomas, Bastian Rieck, Ian Adelstein, and Smita Krishnaswamy. "Diffusion Curvature for Estimating Local Curvature in High Dimensional Data." In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[BMT17] Omer Bobrowski, Sayan Mukherjee, and Jonathan E. Taylor. "Topological consistency via kernel estimation." *Bernoulli*, **23**(1):288–328, 2017.

[Bor48] Karol Borsuk. "On the imbedding of systems of compacta in simplicial complexes." *Fundamenta Mathematicae*, **35**(1):217–234, 1948.

[BSL00] M. Bernstein, V. de Silva, J. C. Langford, and J. B. Tenenbaum. "Graph approximations to geodesics on embedded manifolds." 2000.

[BYM22] Agnese Barbensi, Hee Rhang Yoon, Christian Degnbol Madsen, Deborah O. Ajayi, Michael P.H. Stumpf, and Heather A. Harrington. "Hypergraphs for multiscale cycles in structured data." *arXiv:2210.07545*, 2022.

[BZ18] Andrew Banman and Lori Ziegelmeier. *Mind the Gap: A Study in Global Development Through Persistent Homology*, pp. 125–144. Springer International Publishing, Cham, Switzerland, 2018.

[Car20] G. Carlson. "Topological methods for data modelling." *Nature Reviews Physics*, **2**:697–707, 2020.

[CCL09] F. Chazal, D. Cohen-Steiner, A. Lieutier, and B. Thibert. "Stability of Curvature Measures." *Computer Graphics Forum*, **28**(5):1485–1496, 2009.

[CCM11] Frédéric Chazal, David Cohen-Steiner, and Quentin Mérigot. "Geometric Inference for Probability Measures." *Foundations of Computational Mathematics*, **11**:733–751, 2011.

[CdG16] Frédéric Chazal, Vin de Silva, Marc Glisse, and Steve Oudot. *The Structure and Stability of Persistence Modules*. Springer, Cham, Switzerland, 1st edition, 2016.

[CDM07] Yao-li Chuang, Maria R. D'Orsogna, Daniel Marthaler, Andrea L. Bertozzi, and Lincoln S. Chayes. "State transitions and the continuum limit for a 2D interacting, self-propelled particle system." *Physica D: Nonlinear Phenomena*, **232**(1):33–47, 2007.

[CEF13] Andrea Cerri, Marc Ethier, and Patrizio Frosini. "A Study of Monodromy in the Computation of Multidimensional Persistence." In Rocio Gonzalez-Diaz, Maria-Jose Jimenez, and Belen Medrano, editors, *Discrete Geometry for Computer Imagery*, volume 7749, pp. 192–202, Berlin, Heidelberg, 2013. Springer.

[CEM06] David Cohen-Steiner, Herbert Edelsbrunner, and Dmitriy Morozov. "Vines and Vineyards by Updating Persistence in Linear Time." In *Proceedings of the Twenty-Second Annual Symposium on Computational Geometry*, SCG '06, pp. 119–126, New York, NY, USA, 2006. Association for Computing Machinery.

[Cen] Centers for Disease Control and Prevention. "Risk for COVID-19 Infection, Hospitalization, and Death By Race/Ethnicity." `https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/hospitalization-death-by-race-ethnicity.html` (22 November 2021).

[CF94] Bernard Chazelle and Joel Friedman. "Point location among hyperplanes and unidirectional ray-shooting." *Computational Geometry*, **4**(2):53–62, 1994.

[CG13] Yongwan Chun and Daniel A. Griffith. *Spatial Statistics and Geostatistics: Theory and Applications for Geographic Information Science and Technology*. Sage Publishing, Thousand Oaks, CA, USA, 2013.

[CHM22] Justin Curry, Haibin Hang, Washington Mio, Tom Needham, and Osman Berat Okutan. "Decorated merge trees for persistent topology." *Journal of Applied and Computational Topology*, **6**:371–428, 2022.

[CHP19] M. Keith Chen, Kareem Haggag, Devin G Pope, and Ryne Rohla. "Racial Disparities in Voting Wait Times: Evidence from Smartphone Data." Working Paper 26487, National Bureau of Economic Research, 2019. Available at SSRN: `https://ssrn.com/abstract=3492890`.

[Cit] City of New York. "COVID-19: Data on Vaccines — NYC Health." `https://www1.nyc.gov/site/doh/covid/covid-19-data-vaccines.page` (23 February 2021).

[CJ17] Padraig Corcoran and Christopher B. Jones. "Modelling Topological Features of Swarm Behaviour in Space and Time With Persistence Landscapes." *IEEE Access*, **5**:18534–18544, 2017.

[CMT22] Justin Curry, Sayan Mukherjee, and Katharine Turner. "How many directions determine a shape and other sufficiency results for two topological transforms." *Transactions of the American Mathematical Society, Series B*, **9**:1006–1043, 2022.

[Cor] Lilian Coral. "Los Angeles City Zip Codes." `https://geohub.lacity.org/datasets/lahub::los-angeles-city-zip-codes/about(4April2020)`.

[CS10] Gunnar Carlsson and Vin de Silva. "Zigzag Persistence." *Foundations of Computational Mathematics*, **10**:367–405, 2010.

[CS16] Francesco Camastra and Antonino Staiano. "Intrinsic dimension estimation: Advances and open problems." *Information Sciences*, **328**:26–41, 2016.

[Cur14] Justin Curry. *Sheaves, Cosheaves, and Applications*. PhD thesis, University of Pennsylvania, 2014.

[CZ07] Gunnar Carlsson and Afra Zomorodian. "The Theory of Multidimensional Persistence." *Discrete and Computational Geometry*, **42**:71–93, 2007.

[EH10] Herbert Edelsbrunner and John Harer. *Computational Topology: An Introduction*. American Mathematical Society, Providence, RI, 2010.

[ELZ22] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. "Topological Persistence and Simplification." *Discrete & Computational Geometry*, **28**:511–533, 2022.

[Fac] USA Facts. "US COVID-19 Cases and Deaths by State." https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/ (1 July 2020).

[FHP22] Michelle Feng, Abigail Hickok, and Mason A. Porter. "Topological data analysis of spatial systems." In Federico Battiston and Giovanni Petri, editors, *Higher-Order Systems*, chapter 17, pp. 389–399. Springer, Cham, Switzerland, 2022.

[Fow20] Stephan Fowler. "Why Do Nonwhite Georgia Voters Have To Wait In Line For Hours? Too Few Polling Places." NPR, 2020. Available at https://www.npr.org/2020/10/17/924527679/why-do-nonwhite-georgia-voters-have-to-wait-in-line-for-hours-too-few-polling-pl.

[FP20] M. Feng and M. A. Porter. "Spatial applications of topological data analysis: Cities, snowflakes, random structures, and spiders spinning under the influence." *Physical Review Research*, **2**:033426, 2020.

[FP21] M. Feng and M. A. Porter. "Persistent homology of geospatial data: A case study with voting." *SIAM Review*, **63**(1):67–99, 2021.

[Gap] The Gapminder Foundation. "Gapminder World." http://www.gapminder.com/world (last accessed 29 January 2022).

[Geo] Los Angeles GeoHub. "COVID19 by Neighborhood." https://geohub.lacity.org/datasets/covid19-by-neighborhood/about (3 June 2020).

[GGB16] Chad Giusti, Robert Ghrist, and Danielle S. Bassett. "Two's company, three (or more) is a simplex." *Journal of Computational Neuroscience*, **41**:1–14, 2016.

[Ghr14] Robert Ghrist. *Elementary Applied Topology*. Createspace, 1st edition, 2014.

[GN04] M. T. Gastner and M. E. J. Newman. "Diffusion-based method for producing density-equalizing maps." *Proceedings of the National Academy of Sciences of the United States of America*, **101**(20):7499–7504, 2004.

[Goo] Google Developers. "Distance Matrix API." Available at https://developers.google.com/maps/documentation/distance-matrix. Accessed 4–7 November 2021.

[GS03] James Gimpel and Jason Schuknecht. "Political Participation and Accessibility of the Ballot Box." *Political Geogr.*, **22**:471–488, 2003.

[Hat02] Allen Hatcher. *Algebraic Topology*. Cambridge University Press, Cambridge, UK, 1st edition, 2002.

[HB23] Abigail Hickok and Andrew J. Blumberg. "An intrinsic approach to scalar-curvature estimation for point clouds." *In preparation*, 2023.

[HG19] Jakob Hansen and Robert Ghrist. "Toward a spectral theory of cellular sheaves." *Journal of Applied. and Computational Topology*, **3**:315–358, 2019.

[HGL21] Xiao Hou, Song Gao, Qin Li, Yuhao Kang, Nan Chen, Kaiping Chen, Jinmeng Rao, Jordan S. Ellenberg, and Jonathan A. Patz. "Intracounty modeling of COVID-19 infection with human mobility: Assessing spatial heterogeneity with business traffic, age, and race." *Proceedings of the National Academy of Sciences of the United States of America*, **118**(24):e2020524118, 2021.

[Hic22a] A. Hickok. "Computing Persistence Diagram Bundles." *arXiv:2210.06424*, 2022.

[Hic22b] A. Hickok. "A family of density-scaled filtered complexes." *arXiv:2112.03334*, 2022.

[Hic22c] Abigail Hickok. "Persistence Diagram Bundles: A Multidimensional Generalization of Vineyards." *arXiv:2210.05124*, 2022.

[HJJ22] Abigail Hickok, Benjamin Jarman, Michael Johnson, Jiajie Luo, and Mason A. Porter. "Persistent Homology for Resource Coverage: A Case Study of Access to Polling Sites." *arXiv:2206.04834*, 2022.

[HK05] Moshe Haspel and H. Gibbs Knotts. "Location, Location, Location: Precinct Placement and the Costs of Voting." *J. Politics*, **67**(2):560–573, 2005.

[HK12] Timothy L Hawthorne and Mei-Po Kwan. "Using GIS and perceived distance to understand the unequal geographies of healthcare in lower-income urban neighbourhoods." *Geogr. J.*, **178**(1):18–30, 2012.

[HKF22] A. Hickok, Y. Kureh, M. Feng, H. Z. Brooks, and M. A. Porter. "A bounded-confidence model of opinion dynamics on hypergraphs." *SIAM Journal on Applied Dynamical Systems*, **21**(1):1–32, 2022.

[HNP22] A. Hickok, D. Needell, and M. A Porter. "Analysis of Spatial and Spatiotemporal Anomalies Using Persistent Homology: Case Studies with COVID-19 Data." *SIAM J. Math. Data Sci.*, **4**(3):1116–1144, 2022.

[HSA21] Surendra Hazarie, David Soriano-Paños, Alex Arenas, Jesús Gómez-Gardeñes, and Gourab Ghoshal. "Interplay between population density and mobility in determining the spread of epidemics in cities." *Communications in Physics*, **4**:191, 2021.

[Kan] Jason Kandel. "Timeline: A Look at Key Coronavirus Pandemic Events and Milestones in California." `https://www.nbclosangeles.com/news/coronavirus/2020-2021-california-coronavirus-pandemic-timeline-key-events/2334100/` (15 June 2021).

[Kan19] Henry Kanengiser. "In New York, Where You Live Can Determine How Hard it is to Vote." Available at `https://citylimits.org/2019/04/25/nyc-polling-place-shortage-inequality/`, 2019.

[KLB20] A. Klimovskaia, D. Lopez-Paz, L. Bottou, and M. Nickel. "Poincaré maps for analyzing complex hierarchies in single-cell data." *Nature Communications*, **11**:2966, 2020.

[LB04] Elizaveta Levina and Peter Bickel. "Maximum Likelihood Estimation of Intrinsic Dimension." In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004.

[LD19] Didong Li and David B. Dunson. "Geodesic distance estimation with spherelets." *arxiv:1907.00296*, 2019.

[LRS10] Jesus De Loera, Joerg Rambau, and Francisco Santos. *Triangulations: Structures for Algorithms and Applications*, chapter 6.3.2, pp. 314–316. Springer Science Business Media, 2010.

[LTH21] Lu Li, Connor Thompson, Gregory Henselman-Petrusek, Chad Giusti, and Lori Ziegelmeier. "Minimal Cycle Representatives in Persistent Homology Using Linear Programming: An Empirical Study With User's Guide." *Frontiers in Artificial Intelligence*, **4**, 2021.

[LW15] Michael Lesnick and Matthew Wright. "Interactive Visualization of 2-D Persistence Modules." *arXiv:1512.00180*, 2015.

[LWA17] Yanjie Li, Dingkang Wang, Giorgio A. Ascoli, Partha Mitra, and Yusu Wang. "Metrics for comparing neuronal tree shapes based on persistent homology." *PLoS ONE*, **12**(8):e0182184, 2017.

[LWF03] Xiang-Yang Li, Peng-Jun Wan, and O. Frieder. "Coverage in wireless ad hoc sensor networks." *IEEE Transactions on Computers*, **52**(6):753–763, 2003.

[M21] Woojin Kim Facundo Mémoli. "Spatiotemporal Persistent Homology for Dynamic Metric Spaces." *Discrete & Computational Geometry*, **66**:831–875, 2021.

[Mei93] S. Meiser. "Point Location in Arrangements of Hyperplanes." *Information and Computation*, **106**(2):286–303, 1993.

[MKP01] S. Meguerdichian, F. Koushanfar, M. Potkonjak, and M. B. Srivastava. "Coverage problems in wireless ad-hoc sensor networks." In *Proceedings IEEE INFOCOM 2001. Conference on Computer Communications. Twentieth Annual Joint Conference of the IEEE Computer and Communications Society (Cat. No.01CH37213)*, volume 3, pp. 1380–1387, 2001.

[MTC10] Shawn Martin, Aidan Thompson, Evangelos A. Coutsias, and Jean-Paul Watson. "Topology of cyclo-octane energy landscape." *Journal of Chemical Physics*, **132**(23):234115, 2010.

[MZR16] Slobodan Maletić, Yi Zhao, and Milan Rajković. "Persistent topological features of dynamical systems." *Chaos*, **26**(5):053105, 2016.

[NK17] M. Nickel and D. Kiela. "Poincaré Embeddings for Learning Hierarchical Representations." In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[NYCa] NYC By Natives. "New York City Zip Codes." `https://www.nycbynatives.com/nyc_info/new_york_city_zip_codes.php` (30 March 2021).

[NYCb] NYC Department of Health and Mental Hygiene. "ZCTA vs MODZCTA." `https://github.com/nychealth/coronavirus-data/issues/64` (28 May 2021).

[NYCc] NYC Open Data. "Modified Zip Code Tabulation Areas (MODZCTA)." `https://data.cityofnewyork.us/Health/Modified-Zip-Code-Tabulation-Areas-MODZCTA-/pri4-ifjk/data` (23 February 2021).

[OF03] S. J. Osher and R. Fedkiw. *Level Set Methods and Dynamic Implicit Surfaces*, volume 153. Springer-Verlag, Heidelberg, Germany, 2003.

[OG09] Arkadas Ozakin and Alexander Gray. "Submanifold density estimation." In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.

[OKC20] Alexandra M. Oster, Gloria J. Kang, Amy E. Cha, Vladislav Beresovsky, Charles E. Rose, Gabriel Rainisch, Laura Porter, Eduardo E. Valverde, Elisha B. Peterson, Anne K. Driscoll, Tina Norris, Nana Wilson, Matthew Ritchey, Henry T. Walke, Dale A. Rose, Nadia L. Oussayef, Monica E. Parise, Zack S. Moore, Aaron T. Fleischauer, Margaret A. Honein, Emilio Dirlikov, and Julie Villanueva. "Trends in Number and Distribution of COVID-19 Hotspot Counties — United States, March 8–July 15, 2020." *Morbidity and Mortality Weekly Report*, **69**(33):1127–1132, 2020.

[Ope21] OpenStreetMap contributors. "OpenStreetMap.", 2021. Version 1.1.1. Accessed 19–22 January 2022.

[OPT17] Nina Otter, Mason A. Porter, Ulrike Tillmann, Peter Grindrod, and Heather A. Harrington. "A roadmap for the computation of persistent homology." *EPJ Data Science*, **6**:17, 2017.

[OR87] Joseph O'Rourke. *Art Gallery Theorems and Algorithms*, chapter 5, pp. 125–145. Oxford University Press, Oxford, UK, 1987.

[Pen06] Xavier Pennec. "Intrinsic Statistics on Riemannian Manifolds: Basic Tools for Geometric Measurements." *Journal of Mathematical Imaging and Vision*, **25**:127, 2006.

[Pet06] Peter Petersen. *Riemannian Geometry*, volume 171 of *Graduate Texts in Mathematics*. Springer, New York, NY, 2nd edition, 2006.

[PG16] M. A. Porter and J. P. Gleeson. *Dynamical Systems on Networks: A Tutorial*, volume 4. Springer International Publishing, Cham, Switzerland, 2016.

[Por19] California Open Data Portal. "California County Boundaries." `https://data.ca.gov/dataset/ca-geographic-boundaries/resource/b0007416-a325-4777-9295-368ea6b710e6` (10 September 2019), 2019.

[PWB06] Jamie Pearce, Karen Witten, and Phil Bartie. "Neighbourhoods and health: A GIS approach to measuring community resource accessibility." *J. Epidemiol. Community Health*, **60**(5):389–395, 2006.

[RB19] Raul Rabadan and Andrew J. Blumberg. *Topological Data Analysis for Genomics and Evolution: Topology in Biology*. Cambridge University Press, Cambridge, 2019.

[SBS21] Onkar Sadekar, Mansi Budamagunta, G. J. Sreejith, Sachin Jain, and M. S. Santhanam. "An infectious diseases hazard map for India based on mobility and transportation networks." *Current Science*, **121**(9):1208, 2021.

[SCD22] Kristian Strommen, Matthew Chantry, Joshua Dorrington, and Nina Otter. "A topological perspective on weather regimes." *Climate Dynamics*, 2022.

[Sch21] Rachel Schnalzer. "Traffic is terrible again. Here's how to get it closer to spring 2020 levels." LA Times, 2021. Available at `https://www.latimes.com/business/story/2021-07-22/los-angeles-traffic-congestion-commute-pandemic`.

[SDG18] F. Sala, C. De Sa, A. Gu, and C. Re. "Representation Tradeoffs for Hyperbolic Embeddings." In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4460–4469. PMLR, 2018.

[SG06] Vin de Silva and Robert Ghrist. "Coordinate-free Coverage in Sensor Networks with Controlled Boundaries via Homology." *Int. J. Robot. Res.*, **25**(12):1205–1222, 2006.

[SG07] Vin de Silva and Robert Ghrist. "Coverage in sensor networks via persistent homology." *Alg. Geom. Topology*, **7**(1):339–358, 2007.

[SGR15] Romeil Sandhu, Tryphon Georgiou, Ed Reznik, Liangjia Zhu, Ivan Kolesov, Yasin Senbabaoglu, and Allen Tannenbaum. "Graph Curvature for Differentiating Cancer Networks." *Scientific Reports*, **5**(12323), 2015.

[SHC18] Leo Speidel, Heather A. Harrington, S. Jonathan Chapman, and Mason A. Porter. "Topological data analysis of continuum percolation with disks." *Physical Review E*, **98**:012318, 2018.

[SHP16] Bernadette J. Stolz, Heather A. Harrington, and Mason A. Porter. "The Topological "Shape" of Brexit." *arXiv:1610.00752*, 2016.

[SJS17] R.P. Sreejith, Jürgen Jost, Emil Saucan, and Areejit Samal. "Systematic evaluation of a new combinatorial curvature for complex networks." *Chaos, Solitons & Fractals*, **101**:50–67, 2017.

[SLG20] M. Soliman, V. Lyubchich, and Y. R. Gel. "Ensemble forecasting of the Zika space-time spread with topological data analysis." *Environmetrics*, **31**(7):e2629, 2020.

[SPG19] Ann E. Sizemore, Jennifer E. Phillips-Cremins, Robert Ghrist, and Danielle S. Bassett. "The importance of the whole: Topological data analysis for the network neuroscientist." *Network Neuroscience*, **3**(3):656–673, 2019.

[SSG18] Areejit Samal, R. P. Sreejith, Jiao Gu, Shiping Liu, Emil Saucan, and Jürgen Jost. "Comparative analysis of two discretizations of Ricci curvature for complex networks." *Scientific Reports*, **8**(8650), 2018.

[SSH19] Simon Scarr, Manas Sharma, and Marco Hernandez. "Roads, boats and elephants: How India mobilised a million polling stations." Reuters, 2019. Available at `https://graphics.reuters.com/INDIA-ELECTION-STATIONS/010092FY33Z/index.html`.

[ST86] Neil Sarnak and Robert E. Tarjan. "Planar Point Location Using Persistent Search Trees." *Communications of the ACM*, **29**(7):669–679, 1986.

[STA] STAT. "The Covid-19 Tracker." `https://www.statnews.com/feature/coronavirus/covid-19-tracker/` (16 June 2021).

[SWH21] Duluxan Sritharan, Shu Wang, and Sahand Hormoz. "Computing the Riemannian curvature of image patch and single-cell RNA sequencing data manifolds using extrinsic differential geometry." *Proceedings of the National Academy of Sciences*, **118**(29):e2100473118, 2021.

[SZW21] Ignacio Segovia-Dominguez, Zhiwei Zhen, Rishabh Wagh, Huikyo Lee, and Yulia R. Gel. "TLife-LSTM: Forecasting Future COVID-19 Progression with Topological Signatures of Atmospheric Conditions." In Kamal Karlapalem, Hong Cheng, Naren Ramakrishnan, R. K. Agrawal, P. Krishna Reddy, Jaideep Srivastava, and Tanmoy Chakraborty, editors, *Advances in Knowledge Discovery and Data Mining*, pp. 201–212, Cham, 2021. Springer International Publishing.

[TdL00] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. "A Global Geometric Framework for Nonlinear Dimensionality Reduction." *Science*, **290**(5500):2319–2323, 2000.

[The] The New York Times. "Coronavirus in the U.S.: Latest Map and Case Count." `https://www.nytimes.com/interactive/2021/us/covid-cases.html?action=click&module=Top%20Stories&pgtype=Homepage` (29 January 2022).

[The21] The White House. "FACT SHEET: President Biden Announces 90% of the Adult U.S. Population will be Eligible for Vaccination and 90% will have a Vaccination Site Within 5 Miles of Home by April 19." Available at `https://www.whitehouse.gov/briefing-room/statements-releases/2021/03/29/fact-sheet-president-biden-announces-90-of-the-adult-u-s-population-will-be-eligible-for-vaccination-and-90-will-have-a-vaccination-site-within-5-miles-of-home-by-april-19/`, 2021.

[TKH15] Dane Taylor, Florian Klimm, Heather A. Harrington, Miroslav Kramár, Konstantin Mischaikow, Mason A. Porter, and Peter J. Mucha. "Topological data analysis of contagion maps for examining spreading processes on networks." *Nature Communications*, **6**:7723, 2015.

[TMB14] Katharine Turner, Sayan Mukherjee, and Doug M. Boyer. "Persistent homology transform for modeling shapes and surfaces." *Information and Inference: A Journal of the IMA*, **3**(4):310–344, 2014.

[TMK20] Sarah Tymochko, Elizabeth Munch, and Firas A. Khasawneh. "Using Zigzag Persistent Homology to Detect Hopf Bifurcations in Dynamical Systems." *Algorithms*, **13**(11):278, 2020.

[TOG17] Csaba D. Toth, Joseph O'Rourke, and Jacob E. Goodman. *Handbook of discrete and computational geometry*. CRC Press, Boca Raton, FL, 3rd edition, 2017.

[Tru] Donald Trump. "Proclamation on Declaring a National Emergency Concerning the Novel Coronavirus Disease (COVID-19) Outbreak." `https://trumpwhitehouse.archives.gov/presidential-actions/proclamation-declaring-national-emergency-concerning-novel-coronavirus-disease-covid-19-outbreak/` (13 March 2021).

[TZH15a] C. M. Topaz, L. Ziegelmeier, and T. Halverson. "Topological data analysis of biological aggregation models." *PLOS ONE*, **10**(5):e0126383, 2015.

[TZH15b] Chad M. Topaz, Lori Ziegelmeier, and Tom Halverson. "Topological Data Analysis of Biological Aggregation Models." *PLoS ONE*, **10**(5):e0126383, 2015.

[US] U.S. Census Bureau. "American Community Survey, 2015–2019 Estimates." Available at `https://data.census.gov/`. Accessed 7 November 2021.

[US19] U.S. Census Bureau. "American Community Survey 5-Year Estimates.", 2019.

[VTD20] A. Vespignani, H. Tian, C. Dye, J. O. Lloyd-Smith, R. M. Eggo, M. Shrestha, S. V. Scarpino, B. Gutierrez, M. U. G. Kraemer, J. Wu, K. Leung, and G. M. Leung. "Modelling COVID-19." *Nature Reviews Physics*, **2**(279–281), 2020.

[WAB05] Yusu Wang, Pankaj K. Agarwal, Paul Brown, Herbert Edelsbrunner, and Johannes Rudolph. "Coarse and reliable geometric alignment for protein docking." *Pacific Symposium on Biocomputing*, **10**:64–75, 2005.

[Wei94] Schmuel Weinberger. *The Topological Classification of Stratified Spaces*. Chicago Lectures in Mathematics. The University of Chicago Press, 1994.

[XAT22] Lu Xian, Henry Adams, Chad M. Topaz, and Lori Ziegelmeier. "Capturing Dynamics of Time-Varying Data via Topology." *Foundations of Data Science*, **4**:1–36, 2022.

[YB20] Gökhan Yalnız and Nazmi Burak Budanur. "Inferring symbolic dynamics of chaotic flows from persistence." *Chaos*, **30**(3):033109, 2020.

[ZC05] Afra Zomorodian and Gunnar Carlsson. "Computing Persistent Homology." *Discrete & Computational Geometry*, **33**:249–274, 2005.

[ZH05] Honghai Zhang and Jennifer C. Hou. "Maintaining sensing coverage and connectivity in large sensor networks." *Ad Hoc & Sensor Wireless Netw.*, **1**:89–124, 2005.