Thesis Synopsis

Michael Lesnick*
Institute for Advanced Study

January 30, 2013

My research concerns the theoretical foundations of topological data analysis (TDA). Work done in the last several years on theory and applications of TDA has demonstrated that TDA offers a principled, flexible, and computationally viable framework for studying coarse-scale global geometric features of data [17, 26, 13, 7, 11, 10, 8, 24]. Moreover, the tools of TDA have proven to be very well suited to a wide variety of applications [16, 3, 10, 25, 17, 2, 12, 23].

Nevertheless, in a number of important ways, TDA remains immature as a data analysis methodology. I believe that a more fully developed methodology for TDA, supported by good theory, would greatly broaden the utility and appeal of these tools to statisticians and scientists, and would thus hasten the discovery of applications of TDA to the sciences.

My basic goal as a researcher is to contribute to the theoretical foundations of TDA, with a view towards advancing TDA methodology.

1 Overview of My Thesis Work

My thesis, which I completed in August 2012, focuses on the development of theory of topological inference in the multidimensional persistence setting, where the theory has remained relatively underdeveloped, compared to the 1-D persistence setting. The thesis establishes a number of theoretical results centered around this theme, some of which are interesting even for 1-D persistence. In addition, in a departure from previous work on topological inference, my thesis work presents theory of topological inference formulated directly at the (topological) level of filtrations, rather than only at the (algebraic) level of persistent homology modules.

The main mathematical objects of study in this work are **interleavings**. These are tools for quantifying the similarity between multidimensional filtrations or between multidimensional persistence modules. ϵ-interleavings, the most basic type of interleavings, were introduced for 1-D filtrations and persistence modules in [7] by Chazal et al., where they were used to prove a strong and very useful generalization of the stability of persistence result of [13]; my work generalizes the definition of ϵ-interleavings appearing in [7] in several directions and uses these generalizations to define pseudometrics on multidimensional filtrations and multidimensional persistence modules called **interleaving distances**.

1.1 Theory of Interleavings and Interleaving Distances On Mutidimensional Persistence Modules

The first part of my thesis, adapted from the preprint [21], studies in detail the theory of ϵ-interleavings and interleaving distances on multidimensional persistence modules. I now describe four main results from this part of the thesis.

We will over a fixed field k. We assume that the k-vector spaces in our multidimensional persistence modules are indexed by $\mathbb{R}^n$ (as opposed to being indexed by $\mathbb{N}^n$ or $\mathbb{Z}^n$, as in some work on the theory of persistence).

[*] mlesnick@ias.edu

Let d_I denote the interleaving distance on multidimensional persistence modules, and let d_B denote the usual bottleneck distance on 1-D persistence modules; it follows from the recent structure theorem of William Crawley-Boevey [15] that d_B is well defined on *pointwise finite* persistence modules–that is, on persistence modules whose vector spaces are finite dimensional.

1.1.1 The "Isometry Theorem"

My first main result shows that on pointwise finite 1-D persistence modules, $d_I = d_B$. Together with the structure theorem of [15], this result yields a converse to the main result of [7], answering a question posed in that paper. The result that $d_I = d_B$ has been called the **isometry theorem** by other authors who, since my presentation of the result in the preprint [21], have also provided proofs of the result at varying levels of generality [1],[9].

1.1.2 A Characterization of Interleaved Pairs of Modules

My second main result is a characterization of ϵ-interleaved pairs of persistence modules; it expresses transparently the sense in which ϵ-interleaved persistence modules are algebraically similar. It does this by showing that two multidimensional persistence modules M and N are ϵ-interleaved if and only if there exist presentations[1] of M and N that differ from one another by shifts by $\vec{\epsilon}$ of the grades of generators and relations in the presentations; here $\vec{\epsilon} \in \mathbb{R}^n$ denotes the vector whose components are each ϵ.

This result in turn yields a characterization of d_I.

My characterization of ϵ-interleaved pairs of modules in fact extends, with essentially the same proof, to more general types of interleavings between multidimensional persistence modules which I call (J_1, J_2)-interleavings. This turns out to be important in my study of topological inference later in the thesis.

1.1.3 Optimality of the Interleaving Distance

My third main result is an optimality result for d_I. It tells us that d_I is stable in a sense analogous to that which d_B is shown to be stable in [13, 7], and further that when the underlying field k is $\mathbb{Q}$ or a field of prime order, if d is another stable pseudometric on multi-dimensional persistence modules then $d \leq d_I$.

Motivation for the Optimality Result A number of papers have introduced pseudometrics on multidimensional persistence modules [20, 4, 6, 18], and several of these [4, 6, 18] have presented stability results for the metrics they introduce. The multi-dimensional matching distance of [6] is unique amongst the pseudometrics introduced by these papers in that it is a generalization of d_B.

However, in choosing a multidimensional generalization of d_B for use in the development of theory or in applications, we want more of our distance than just good stability properties. A stability result of the kind typically appearing in the persistent homology literature [13, 7, 14, 6] tells us that our distance on persistence modules is not, in some relative sense, too sensitive. However, a good choice of distance should also not be too insensitive; as an extreme example of this consider the pseudometric on persistence modules which is identically 0; it satisfies lots of strong stability properties, yet is clearly too insensitive to be of any use.

Ideally, then, we would like to have a generalization of d_B on multidimensional persistence modules which is not only stable, but also is as sensitive as a stable metric can be, in a suitable sense. My optimality result shows that d_I satisfies a property of this kind, and that it is the unique distance which does so. The result thus distinguishes d_I from the many possible choices of stable distances on multidimensional persistence modules as a particularly natural choice for use in the development of theory.

[1] For the definition of a presentation of a multidimensional persistence module see [21] or [22].

1.1.4 A Result on the Computation of the Interleaving Distance

My fourth main result is that given presentations for two finitely presented multidimensional persistence modules M and N, the problem of computating $d_I(M,N)$ is polynomially equivalent to deciding the solvability of $O(\log m)$ systems of multivariate quadratic equations, each with $O(m^2)$ variables and $O(m^2)$ equations, where m is the total number of generators and relations in the presentations for M and N.

1.2 Interleavings and Interleaving Distances on Multidimensional Filtrations and Applications to Topological Inference

In the second part of the thesis we define interleavings and interleaving distances on multidimensional filtrations, and we present a theoretical treatment of these parallel to the treatment of interleavings and interleaving distances on multidimensional persistence modules in the first part of the thesis. We then use interleavings and interleaving distances on multidimensional filtrations to formulate and prove 2-D adaptations of a (loose) analogue for persistent homology of the weak law of large numbers.[2] We formulate these "weak laws for 2-D persistence" directly on the level of filtrations.

To do this, we employ localization of categories, a standard construction in homotopy theory, to define and study homotopy theoretic versions of interleavings and the interleaving distance on multidimensional filtrations, which we call **homotopy interleavings** and the **homotopy interleaving distance** d_{HI}**.**[3] We formulate our weak laws for 2-D persistence using homotopy interleavings and d_{HI}.

Our weak laws on the level of filtrations yield as corollaries analogous results on the level of persistent homology, formulated in terms of the interleaving distance and interleavings on persistence modules. Our characterization of (J_1,J_2)-interleavings from the first part of our thesis yields concrete interpretations of these corollaries as statements about the similarity between presentations of persistent homology modules.

We will now describe our weak laws for 2-D persistence in more detail.

1.2.1 Notation and Terminology

Fix a field k and $i \in \mathbb{Z}_{\geq 0}$.

For $T \subset \mathbb{R}^m$, and $b \in \mathbb{R}$, let $\check{\text{C}}\text{ech}(T,b)$ denote the Čech complex on T with scale parameter b, and let $\text{Rips}(T,b)$ denote the Vietoris-Rips complex on T with scale parameter b.

Let's agree to regard $\mathbb{R}$, together with its usual total order $<$, as a category in the usual way. $\mathbb{R}^{op} \times \mathbb{R}$ is then a poset category; we denote the partial order relation by $<$. For C a category, Let $C^{\mathbb{R}^{op}\times\mathbb{R}}$ denote the category of diagrams of objects in C indexed by $\mathbb{R}^{op} \times \mathbb{R}$. We call objects in $\mathbf{Top}^{\mathbb{R}^{op}\times\mathbb{R}}$ 2-filtrations and objects in $\mathbf{Vect}_k^{\mathbb{R}^{op}\times\mathbb{R}}$ 2-persistence modules. For $X \in \text{obj}(C^{\mathbb{R}^{op}\times\mathbb{R}})$ and $a \in \mathbb{R}^2$, we let X_a denote the object of C in X at index a; for $a \leq b \in \text{obj}(\mathbb{R}^{op} \times \mathbb{R})$, we let $\varphi_X(a,b)$ denote the unique morphism in X from a to b.

The singular homology functor with coefficients in k induces a functor $H_i : \mathbf{Top}^{\mathbb{R}^{op}\times\mathbb{R}} \to \mathbf{Vect}_k^{\mathbb{R}^{op}\times\mathbb{R}}$, which we call the i^{th} **persistent homology functor**.

Let $\gamma : \mathbb{R}^m \to \mathbb{R}$ be a Lipchitz probability density function; for $z \in \mathbb{N}$, let T_z be an i.i.d. sample of a probability distribution with density γ.

Let E be a density estimator[4] with the property that $E(T_z)$ converges uniformly in probability to γ as $z \to \infty$. This condition on E satisfied, for example, by kernel density estimators, under mild conditions on the kernel and on γ, provided the kernel bandwidth tends to 0 at the appropriate rate as z increases [19].

[2] The "weak law for persistent homology" which we adapt to the multidimensional setting is implicit in the work of Chazal, Guibas, Oudot, Skraba [10] and is certainly of interest in its own right. See my thesis [22, Chapter 1] for a discussion of this.

[3] In my thesis, I called these "weak interleavings" and "the weak interleaving distance," but I have decided that it is better to replace the word "weak" with "homotopy" in these names.

[4] Here, a density estimator E is a simply function which associates to each finite set of points $L \subset \mathbb{R}^m$ a probability density $E(L) : \mathbb{R}^m \to \mathbb{R}$.

1.2.2 Three 2-Filtrations

We define three types of 2-filtrations. Let U denote a 1-point topological space. In each of our filtrations X, the maps $\varphi_X(a,b)$ will be defined in the obvious way either as inclusions or as (the unique) maps to U; we omit the explicit specification of these maps.

Define $F^{SO}(\gamma) \in \text{obj}(\mathbf{Top}^{\mathbb{R}^{op}\times\mathbb{R}})$ by taking

$$F^{SO}(\gamma)_{(a,b)} = \begin{cases} \{y \in \mathbb{R}^m | d(\gamma^{-1}([a,\infty)), y) \leq b\} & \text{if } a \geq 0 \\ U & \text{otherwise.} \end{cases}$$

For T any finite subset of $\mathbb{R}^m$, define $F^{\text{SČe}}(T,E),\ F^{SR}(T,E) \in \text{obj}(\mathbf{Top}^{\mathbb{R}^{op}\times\mathbb{R}})$ by taking

$$F^{\text{SČe}}(T,E)_{(a,b)} = \begin{cases} \text{Čech}(T \cap E(T)^{-1}([a,\infty)), b) & \text{if } a \geq 0 \\ U & \text{otherwise.} \end{cases}$$

$$F^{SR}(T,E)_{(a,b)} = \begin{cases} \text{Rips}(T \cap E(T)^{-1}([a,\infty)), b) & \text{if } a \geq 0 \\ U & \text{otherwise.} \end{cases}$$

1.2.3 Statement of Results

If $\{X_z\}_{z\in\mathbb{N}}$ is a sequence of random variables and X is a random variable such that $\{X_z\}_{z\in\mathbb{N}}$ converges in probability to X, we write $X_z \xrightarrow{\mathcal{P}} X$.

Theorem 1.1. $d_{HI}(F^{\text{SČe}}(T_z,E), F^{SO}(\gamma)) \xrightarrow{\mathcal{P}} 0.$

Informally, this theorem tells us that in the $z \to \infty$ limit, $F^{\text{SČe}}(T_z,E)$ and $F^{SO}(\gamma)$ are topologically equivalent, in a suitable sense.

The result descends to the level of persistent homology:

Corollary 1.2. $d_I(H_i(F^{\text{SČe}}(T_z,E)), H_i(F^{SO}(\gamma))) \xrightarrow{\mathcal{P}} 0.$

We can also state analogues of the above results for the filtrations $F^{SR}(T_z,\gamma)$. To state such an analogue directly at the level of 2-filtrations requires the language of homotopy (J_1,J_2)-interleavings, introduced in my thesis; we will not state the result here. To precisely state a version of the result at the level of 2-persistence modules also requires more formalism than I wish to present in this research statement; however, we can easily give an informal statement of the result which captures the key idea:

Theorem (Informally Stated) 1.3. *In the $z \to \infty$ limit, there exists a presentation $\langle G|R\rangle$ for $H_i(F^{SR}(T_z,E))$ such that if we multiply the second coordinates of the grades of some of the elements of G and R by $\sqrt{2}$, we obtain a presentation for $H_i(F^{SO}(\gamma))$.*

1.2.4 Remarks on Theorem 1.3

In the original paper on multidimensional persistence [5], Carlsson and Zomorodian proposed use of (a slight variant) of the 2-persistence module $H_i(F^{SR}(T_z,E))$ to study features of the data T_z in a way that is robust to low density noise.

Implicit in that proposal was the the idea that $H_i(F^{SR}(T_z,E))$ should encode topological information about the probability density γ which generated the data T_z. To my knowledge, Theorem 1.3 is the first result giving formal expression to this idea.

References

[1] P. Bubenik and J.A. Scott. Categorification of persistent homology. *arXiv preprint arXiv:1205.3669*, 2012.

[2] E. Carlsson, G. Carlsson, V. De Silva, and SJ Fortune. An algebraic topological method for feature identification. *International Journal of Computational Geometry and Applications*, 16(4):291–314, 2006.

[3] G. Carlsson, T. Ishkhanov, V. De Silva, and A. Zomorodian. On the local behavior of spaces of natural images. *International Journal of Computer Vision*, 76(1):1–12, 2008.

[4] G. Carlsson and F. Mémoli. Multiparameter hierarchical clustering methods. *Classification as a Tool for Research*, pages 63–70, 2010.

[5] G. Carlsson and A. Zomorodian. The theory of multidimensional persistence. *Discrete and Computational Geometry*, 42(1):71–93, 2009.

[6] A. Cerri, B. Di Fabio, M. Ferri, P. Frosini, and C. Landi. Multidimensional persistent homology is stable. *Arxiv preprint arXiv:0908.0064*, 2009.

[7] F. Chazal, D. Cohen-Steiner, M. Glisse, L.J. Guibas, and S.Y. Oudot. Proximity of persistence modules and their diagrams. In *Proceedings of the 25th annual symposium on Computational geometry*, pages 237–246. ACM, 2009.

[8] F. Chazal, D. Cohen-Steiner, L.J. Guibas, F. Mémoli, and S.Y. Oudot. Gromov-Hausdorff stable signatures for shapes using persistence. In *Proceedings of the Symposium on Geometry Processing*, pages 1393–1403. Eurographics Association, 2009.

[9] F. Chazal, V. de Silva, M. Glisse, and S. Oudot. The structure and stability of persistence modules. *arXiv preprint arXiv:1207.3674*, 2012.

[10] F. Chazal, L.J. Guibas, S. Oudot, and P. Skraba. Persistence-based clustering in Riemannian manifolds. *INRIA Technical Report*, 2009.

[11] F. Chazal, L.J. Guibas, S.Y. Oudot, and P. Skraba. Analysis of scalar fields over point cloud data. In *Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1021–1030. Society for Industrial and Applied Mathematics, 2009.

[12] C. Chen and D. Freedman. Measuring and computing natural generators for homology groups. *Computational Geometry*, 43(2):169–181, 2010.

[13] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. *Discrete and Computational Geometry*, 37(1):103–120, 2007.

[14] D. Cohen-Steiner, H. Edelsbrunner, J. Harer, and Y. Mileyko. Lipschitz functions have L p-stable persistence. *Foundations of Computational Mathematics*, 10(2):127–139, 2010.

[15] W. Crawley-Boevey. Decomposition of pointwise finite-dimensional persistence modules. *arXiv preprint arXiv:1210.0819*, 2012.

[16] V. de Silva and R. Ghrist. Coverage in sensor networks via persistent homology. *Algebraic and Geometric Topology*, 7:339–358, 2007.

[17] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Discrete and Computational Geometry*, 28(4):511–533, 2002.

[18] P. Frosini. Stable comparison of multidimensional persistent homology groups with torsion. *Acta Applicandae Mathematicae*, pages 1–12, 2010.

[19] E. Giné and A. Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 38(6):907–921, 2002.

[20] T. Ishkhanov. A topological method for shape comparison. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, pages 1–4. IEEE, 2008.

[21] M. Lesnick. The optimality of the interleaving distance on multidimensional persistence modules. *Arxiv preprint arXiv:1106.5305*, 2011.

[22] M. Lesnick. Multidimensional interleavings and applications to topological inference. *Ph.D. Dissertation, Stanford University*, 2012.

[23] M. Nicolau, A.J. Levine, and G. Carlsson. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences*, 108(17):7265–7270, 2011.

[24] G. Singh, F. Mémoli, and G. Carlsson. Topological methods for the analysis of high dimensional data sets and 3d object recognition. In *Eurographics Symposium on Point-Based Graphics*, volume 22. The Eurographics Association, 2007.

[25] P. Skraba, M. Ovsjanikov, F. Chazal, and L. Guibas. Persistence-based segmentation of deformable shapes. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 45–52. IEEE, 2010.

[26] A. Zomorodian and G. Carlsson. Computing persistent homology. *Discrete and Computational Geometry*, 33(2):249–274, 2005.