Generalized persistence and applications

organized by
Anthony Bak and Dmitriy Morozov

Workshop Summary

Several themes emerged during the workshop.

Factoring through sets. A standard way of encoding multi persistence features of data sets is via functors indexed by the poset $\mathbb{N}^r$ with values in vector spaces. Unfortunately the moduli of such functors, for $r > 1$ and a given dimension table, is a very complicated algebraic variety (this is in contrast with $r = 1$, where this moduli is discrete parametrized by the bar codes). For the purpose of applications such functors are therefore not very useful particularly since functors that come from applications are rather special. Many of them can be, for example, described as functors with values not in vector spaces but sets. Such functors with a given dimension table can be parametrized by discrete invariants. One objective is to identify those that are adapted for calculations. One should however not strive to describe them all. That is still an impossibly difficult task. Instead we should built a hierarchy of them that capture more and more information with higher and higher cost. One strategy is to find sub-quivers of $\mathbb{N}^r$ whose either set representations we understand or at least that have certain known canonical representations. Sub-quivers given by n-arrows starting in one node and ending in n different nodes could be useful for this purpose. A typical irreducible representation is given by the identities of the one dimensional vector space for some of the arrows, with the rest having the trivial vector space as the range. Some discrete invariants can be obtained, for example, as the number of summands of this representation in the restriction the functor to various such sub-quivers. Representations of these quivers also play an important role in glueing the functor indexed by the entire grid $\mathbb{N}^r$ from functors indexed by various sub-cubes and smaller sub-grids of $\mathbb{N}^r$, encoding the glueing data. Having set valued functors is essential here. One could ask how restrictive this assumption is? The philosophy is that by increasing the parameter r, we can convert relevant vector valued to set valued functors. For example, if we are interested in functors whose values are the n-th homology of a r-multifiltration of a space, then its persistence features are expected to be recovered from the persistence features of the 0-th homology of a $(r + n)$-multifiltration.

Generalized persistence and sheaves. Persistent homology seeks to answer the question "Given a family of topological spaces, what homological features develop, dissipate, persist?" For the sheaf-theoretic perspective, a family of spaces $\{X_s\}_{s\in S}$ is best understood as the fibers of a map $f : X \to S$. One way of addressing this question is to observe that there is a natural functor from the (opposite) category of open sets and their inclusions to the category of vector spaces and linear maps by assigning to an open set $U \subset S$ the vector space $F^i(U) = H^i(f^{-1}(U))$. This defines a pre-sheaf, but there is a naturally associated sheaf $\mathcal{F}^i$ that has better properties. The characteristic property that the sheaf $\mathcal{F}^i$ has that the pre-sheaf F^i doesn't have is the ability to go from local to global. For applications to persistence this manifests in two ways: 1) Indecomposables live directly on the parameter space, in

contrast to the collection of open sets of parameter space, thus whereas one needs diamonds in 1D level-set pre-sheaf persistence, intervals suffice for 1D level-set sheaf persistence. 2) One can parallelize homology computations and then use a sheaf to aggregate the results as a sort of database of features, thus if one interrogates the homology over a large region of parameter space, one can use already performed computations to do so.

There are other, more sociological, reasons why sheaves should hold a special place in generalized persistence. There is an extant literature that has been developed by some of the greatest minds of the 20th century, notably Grothendieck, who had an unmatched talent for identifying structure. Also, the development of the interleaving distance for persistence has spawned analogous metrics for the category of sheaves. This could foster remarkable discoveries in pure mathematics, much as the Gromov-Hausdorff distance did. Finally, the reformulation of persistence in terms of sheaves is a necessary first step before jumping into the remarkable developments of Bob MacPherson and Amit Patel, who have identified a 2-categorical analog of the above sheaf that tracks *stable* homological features over a parameter space S, at least when S is a manifold.

One of the problems discussed at the workshop asks whether there is a simpler description of the construction of MacPherson and Patel. Their construction defines the persistent homology class over an open set $U \subset S$ to be the image of the map $H^{n+i}(X, X - f^{-1}(U)) \to H_i(f^{-1}(U))$, which comes from taking the cap product with the pullback of the fundamental class of S. The 2-category structure emerges from connecting these images across multiple open sets. The alternative candidate considered at the workshop also considers the image of a suitably defined map, but with the added advantage that S does not need to be a manifold. Here one considers the image from $H^i(f^{-1}(U))$ to the colimit $\varinjlim_{U \supset V \neq \varnothing} H^i(f^{-1}(V))$. In sheaf-theoretic language this image picks out sections that are supported at every point in U. When $S = \mathbb{R}$, we convinced ourselves that these constructions agree. However, when $S = \mathbb{R}^2$ they differ in a key counterexample: Consider the annulus $X = S^1 \times [0, 1]$ and the map that carries $S^1 \times 1$ to the boundary of the unit disc and carries $S^1 \times 0$ to the origin. Here for $i = 0$, the MacPherson-Patel construction assigns 0 to the origin, but the "persistent sections" construction gives the ground field. The MacPherson-Patel answer is arguably the right one since a small perturbation could "open up the hole" and make the fiber over the origin empty.

From Morse functions to Whitney stratifiable maps. A Morse function is a kind of generic smooth function $f : M \to \mathbb{R}$ with many tameness properties. An important tameness property is the following. If $\{v_0, \ldots, v_n\}$ are the set of critical values of f, then f is a fiber bundle over each open interval (v_i, v_{i+1}). Another class of functions with this tameness property is the class of piecewise linear maps $g : |K| \to \mathbb{R}$. If $\{u_0, \ldots, u_m\}$ are the images of the vertices of f, then f is a fiber bundle over each open interval (u_i, u_{i+1}).

The class of Whitney stratifiable maps contains both these classes of functions plus more. A Whitney stratified space (M, S) is a manifold M decomposed into manifold pieces S called strata that come together nicely. Mark Goresky showed that every Whitney stratified space admits a triangulation. A Whitney stratified map $f : (N, R) \to (M, S)$ is a proper continuous map $f : N \to M$ that takes strata to strata. Rene Thom showed that every Whitney stratified map $f : (N, R) \to (M, S)$ is a fiber bundle over each stratum of S.

Amit Patel described the persistent homology group for Whitney stratifiable maps $f : (N, R) \to (M, S)$ to an oriented m-manifold M. Oriented means $H^m(M)$, cohomology with compact support, is isomorphic to the copy of the field and a generator is chosen. Let

U be a connected open set of M. Then there is a map $H_{d+m}(N, N - f^{-1}(U)) \to H_d(f^{-1}(U))$ defined as the cap product with the pullback of the orientation. The image of this map is the persistent homology over U. This is a generalization of the persistent homology group for functions. Furthermore, the cap product is a homotopy invariant and therefore the persistent homology group encodes stable information.

Persistent cohomology ring. During the workshop, participants constructed a cup product structure on the persistent cohomology of a filtered space. This operation provides another invariant of a filtration that may be useful. We constructed an example illustrating that the persistent cup product can sometimes distinguish between different filtrations having the same persistent homology.

The core idea is that the cup product can be interpreted as a graded linear map from the 2-fold tensor product of cohomology to the cohomology of a space: $H^*(X, k) \otimes H^*(X, k) \to H^*(X, k)$, where X is a topological, space, k is a field, and H^* denotes cohomology.

Since the cup product is functorial, applying this concept to a filtration of topological spaces, on the level of the individual spaces, gives rise to a morphism of persistence modules. We can now consider the image of this morphism, which we may call the persistent cup product module. This is a submodule of the persistent cohomology of the filtration. In particular, it has a barcode, which provides information beyond the persistent (co)homology of the filtration.

Persistence space. A major issue in multidimensional persistence is that, when filtrations depend on multiple parameters, it is not possible to provide complete and discrete representations for multidimensional persistence modules analogous to that provided by persistence diagrams for one-dimensional persistence modules. This theoretical obstruction discouraged so far the introduction of a multidimensional analogue of the persistence diagram. Given the importance of persistence diagrams for the use of persistence in concrete tasks, one can immediately see that the lack of such an analogue is a severe drawback for the actual application of multidimensional persistence. Therefore a natural question we may ask is the following one: In which other sense may we hope to construct a generalization of a persistence diagram for the multidimensional setting?

In order to answer this question, it is useful to observe that the persistence diagram is known to satisfy the following important properties:

- it can be defined via multiplicities obtained from persistent Betti numbers;
- it allows to completely reconstruct persistent Betti numbers;
- the coordinates of its off-diagonal points are homological critical values.

Therefore, it is reasonable to require that a generalization of a persistence diagram for the multidimensional setting satisfies all these properties. It is worth to underlying that because of the aforementioned impossibility result, no generalization of a persistence diagram exists that can achieve the goal of representing completely a persistence module, but only its persistent Betti numbers.

Along these lines, the notion of a persistence diagram can be generalized as follows. A persistence space can be defined as a multiset of points defined via multiplicities so that in the one-dimensional case it coincides with persistence diagrams. Moreover, it allows for a complete reconstruction of multidimensional persistent Betti numbers, and the coordinates of its off-diagonal points are multidimensional homological critical values. Last but not least, the persistent space is stable with respect to the Hausdorff distance.

Computable invariants for multidimensional persistence. Martina Scolamiero presented an algorithmic approach to the description of multidimensional persistence modules of a point cloud. She described an algorithm to compute a presentation by generators and relations of multidimensional persistence modules. The key point in the algorithm was to observe that multidimensional persistence modules are constructed from the modules of n-chains of a multi-filtration and such modules have a special structure and can be decomposed as direct sum of monomial ideals. Finally some discrete but not complete invariants were introduced, the so called Betti tables. Such invariants can be computed very efficiently because of their local nature. How informative the Betti tables are about the shape of a point cloud, once a type of multi-filtration is fixed, and questions about the stability of such invariants were proposed and discussed throughout the workshop.

Missing/Incomplete data. Although working with missing or non complete data is rather classical in many data analysis areas, to our knowledge there does not exist any proven method allowing to analyze data from a topological perspective when some observations are missing or partly missing. The goal of this workgroup was to discuss a few directions that could be explored to address this problem in TDA.

We first discussed to versions of the problem: (i) the case where the data are (possibly) not embedded in an Euclidean space and just come as a matrix of pairwise distances between the data points, but some entries of the matrix are missing. (ii) the case where the data are represented by point clouds in Euclidean spaces, but some of the coordinates of some of the points are missing.

the first case has been quickly discussed but we concentrated our efforts on the second case. We started with a simple version of the problem: - let $P = \{p_1, \cdots, p_n\}$ be a set of points in $\mathbb{R}^D$ such that for each $p_i = (p_{i,1}, \cdots, p_{i,D})$, at most one of the coordinates $p_{i,j}$ is unknown. We also assumed that the points of P are (densely) sampled on a compact d-dimensional manifold M. The problem is then to recover relevant topological information about M (e.g. its betti numbers).

After a few discussions we started to explore the following idea. The point set P with missing coordinates information can be seen as a set of lines in $\mathbb{R}^D$ parallel to the coordinate axes. It is then appealing to try to recover the position of the points of P along these lines in order to reconstruct a (dense) sample of M to which standard TDA methods could be applied. Now, if two points in P are close to each other, then the smallest distance between the two corresponding lines has to be small too. So the idea was to create a new point cloud according to a procedure associating a new point for each pair of lines whose minimal distance is below some threshold (e.g. the middle of the segment joining the two closest points on the lines). Obviously such a procedure create some false positives, i.e. points that appear away from M. However, it seems (at least in some cases) that the density of created points is higher around M, so some density estimate or other methods could be used to filter out the false negative. We run a few very simple experiments that were encouraging.

Computing Interleaving Distance. Our workgroup studied the problem of computing the interleaving distance between multidimensional persistence modules. For this, it suffices to find an efficient algorithm to decide, for any $d \geq 0$, whether two given persistence modules are d-interleaved.

There are known algorithms in the computational algebra literature for deciding whether two persistence modules are 0-interleaved, i.e. isomorphic. One approach is presented in the

paper "Testing isomorphism of modules" Peter A. Brooksbank Eugene M. Luks. Our workgroup studied this paper and attempted to extend its approach to the case of general $d \geq 0$. Our impression, following a couple of days of work and some subsequent thought, is that approach of the paper does not extend readily.

Subsequently, some of the participants of the workgroup have explored a couple of other lines of attack, which also have not yet borne fruit. Thus, the problem of if and how the multi-D interleaving distance can be computed remains open, and continues to be an intriguing challenge.

It is worth noting that while we did not solve the problem, our discussions at AIM about the problem of computing the interleaving distance led to a short but valuable paper on a closely related topic by Claudia Landi, "The rank invariant stability via interleavings." http://arxiv.org/abs/1412.3374.