APPLIED TOPOLOGY LECTURE NOTES

CHAD GIUSTI

UPDATED: NOVEMBER 6, 2017

Topology of Data

When we move to the realm of science and engineering one of the fundamental changes from that of mathematics is that we only get to measure things, rather than assuming we know what we're studying – to put it another way, we move from topological spaces to a finite sample of points from those spaces. However, with a bit of cleverness, we can still use our topological tools to recover information about the underlying spaces in a sensible way.

Barcodes and persistence diagrams

Let's investigate this last collection of terminology a bit. Cycles are elements of a vector space, which are at some point *born* and then *die*. That these are discrete events that apply to each cycle suggests that we should be able to understand the persistent homology of a filtered complex through such a lens.

Fix $[x] \in H_p(S_{t(i)})$, $[x] \neq [0]$ born at $t(i) = b_{[x]}$ and die at $t(j) = d_{[x]}$. Write $[x_{t(k)}] = (\iota_{t(i),t(k)})_*([x]) \in H_p(S_{t(k)})$ and consider the maximal sequence $([x_{t(i)}], [x_{t(i+1)}], \dots [x_{t(j)}])$ of non-zero cycles in the image of $[x_{t(i)}]$ under the induced maps in homology. Without loss of generality, assume that $b_{[x]}$ is minimal among all possible choices of $[x]$.

Since each of these is a non-zero element, we can consider their spans as a sequence of subspaces of the appropriate homology groups

$$\langle [x_{t(k)}] \rangle = \mathbb{F}_2 \langle [x_{t(k)}] \rangle \subseteq H_p(S_{t(k)}).$$

By definition, the restriction

$$(\iota_{t(k),t(k+1)})_p|_{\langle [x_{t(k)}] \rangle} : \langle [x_{t(k)}] \rangle \to \langle [x_{t(k+1)}] \rangle$$

has matrix (1). Without loss of generality, we can assume that all other basis vectors for each homology group are contained in the orthogonal subspaces $\langle [x_{t(k)}] \rangle^{\perp}$, so the persistent homology decom-

poses as

$$\begin{array}{ccccccccccccc}
\cdots & \xrightarrow{0} & 0 & \xrightarrow{0} & \langle [x_{t(i)}] \rangle & \xrightarrow{1} & \langle [x_{t(i+1)}] \rangle & \xrightarrow{1} \cdots \xrightarrow{1} & \langle [x_{t(j)}] \rangle & \xrightarrow{0} & 0 & \xrightarrow{0} & \cdots \\
 & & \oplus & & \oplus & & \oplus & & \oplus & & \oplus & & \\
\cdots & \longrightarrow & H_p(S_{t(i-1)}) & \longrightarrow & \langle [x_{t(i)}] \rangle^{\perp} & \longrightarrow & \langle [x_{t(i+1)}] \rangle^{\perp} & \longrightarrow \cdots \longrightarrow & \langle [x_{t(j)}] \rangle^{\perp} & \longrightarrow & H_p(S_{t(j+1)}) & \longrightarrow & \cdots
\end{array}$$

Write $\text{Int}_{b_{[x]}, d_{[x]}}$ for the *($(b_{[x]}, d_{[x]})$-interval module* given by the sequence of 1-dimensional $\mathbb{F}_2$ vector spaces indexed by $t(i), t(i+1), \ldots, t(j)$ with identity maps between consecutive spaces, as in the first row of the diagram above, and 0-dimensional vector spaces outside that range.

We can inductively apply this argument to the second row of the diagram to split off a sequence of interval modules[1]. Because each homology group is finite dimensional, this process will eventually terminate, resulting in a decomposition of the form

$$H_p(S) \cong \bigoplus_{a=1}^{M} \text{Int}_{b_{[x_a]}, d_{[x_a]}}.$$

[1] This is directly analogous to the Gram-Schmidt process for constructing an orthonormal basis for a vector space, and is a special case of the classification of finitely generated modules over a PID.

Thus, the degree p persistent homology of a filtered simplicial complex can be summarized by a collection of (possibly repeated) pairs, $((b_{[x_a]}, d_{[x_a]}))_{a=1}^{M}$, of birth and death indices for a collection of *generating cycles* $GC(S) = \{[x_a]\}_{a=1}^{M}$ in $H_p(S)$.

There are two common ways to visualize this data, *barcodes* and *peristence diagrams*. Both are useful for understanding the structure of persistent homology, so we will present and use both.

Definition. Let S be a filtered simplicial complex with $H_p(S) \cong \bigoplus_{a=1}^{M} \text{Int}_{b_{[x_a]}, d_{[x_a]}}$. The *degree p persistence diagram for S* is $D(S) = (\text{im}(P), \mu)$, where the function $P : GC(S) \to \mathbb{R}^2$ is given by $P([x_a]) = (b_{[x_a]}, d_{[x_a]})$, and the multiplicity function $\mu : \text{im(P)} \to \mathbb{N}_{>0}$ is given by $\mu((b,d)) = |P^{-1}((b,d))|$.

We draw a persistence diagram by plotting the image and labeling points with their multiplicity. Generically, each pair $(b_{[x_a]}, d_{[x_a]})$ is distinct, so we suppress any label of 1. Also, it is necessarily the case that $0 < b_{[x_a]} < d_{[x_a]} < \infty$, so these points are contained in the upper half of the first quadrant.

Definition. Let S be a filtered simplicial complex with $H_p(S) \cong \bigoplus_{a=1}^{M} \text{Int}_{b_{[x_a]}, d_{[x_a]}}$. The *degree p barcode for S* is the collection of line segments joining the points $(b_{[x_a]}, \frac{1}{a})$ and $(d_{[x_a]}, \frac{1}{a})$, $a = 1, \ldots, M$.

It is common to sort the intervals to make the barcode visually clearer, either in terms of increasing birth or death times.

Distance and stability for persistence

Suppose we compute persistent homology for a pair of filtrations. How do we know if they're "similar"? Given our ability to summarize the structure using diagrams or barcodes, the popular answer is as one might expect: assign a notion of distance to the persistence diagrams.

Naively, we would like to compute this distance by building on some notion of distance between paired points, for which we would need a bijection (with multiplicity) of the points in the two diagrams. However, there is no reason to assume that there are the same number of points. To fix this, we note that in a persistence diagram, points (b, d) for which $\ell = d - b$ is small are "close to vanishing" – they could reasonably be deleted, because the corresponding cycle only appears for a short while in the persistent homology. Thus, we can reasonably pair points with the nearest point on the line $b = d$.

Definition. Let $D_i = (\mathrm{im}(P_i), \mu_i)$, $i = 1, 2$, with $P_i : GC_i \to \mathbb{R}^2$ be persistence diagrams. A *partial matching* for D_1 and D_2 is a bijection $\phi : M_1(\phi) \to M_2(\phi)$ for some choice of subsets $M_i(\phi) \subseteq GC_i$.

A partial matching tells us which points to pair with points from the other diagram; those unmatched in either diagram will be matched to the nearest points on the $(b = d)$ line.

We will use the so-called ∞-norm to measure distance: the distance between two points is the largest amount they differ in a single coordinate. That is, $||(b, d)||_\infty = \max(|b|, |d|)$. In this context, the ∞-norm of a difference of points measures larger of the discrepancies between the birth and death time of the two cycles, which is perhaps more natural than a Euclidean norm-style combination of the two.

Definition. Let D_1 and D_2 be persistence diagrams. Let $\phi : M_1(\phi) \to M_2(\phi)$ be a partial matching for D_1 and D_2, and define

$$d_\phi(D_1, D_2) = \max\{||P_1([x_a]) - P_2(\phi([x_a]))||_\infty \mid [x_a] \in M_1(\phi)\}$$

and

$$d_{\phi^C}(D_1, D_2) = \max\{\ell_{[x_a]} \mid [x_a] \in (GC_1 \setminus M_1(\phi)) \cup (GC_2 \setminus M_2(\phi))\}.$$

The *bottleneck distance* between D_1 and D_2 is

$$d_B(D_1, D_2) = \min_{\phi : M_1(\phi) \to M_2(\phi)} \max\{d_\phi(D_1, D_2), d_{\phi^C}(D_1, D_2)\}.$$

Here, d_ϕ is the largest difference in the birth or death time of a matched generating cycle, and d_{ϕ^C} is the largest lifetime of an unmatched cycle (or, equivalently, the difference in birth or death to

the nearest point on the $b = d$ line). The bottleneck distance arises from selecting the matching with the smallest such discrepancy.

The bottleneck distance has one fundamentally nice property that makes it useful in the context of data analysis: perturbing the input data slightly only perturbs the diagram slightly in terms of d_B. First, we need the "right" notion of distance between point clouds, so we know what a small perturbation is.

Definition. Let $P = \{p_i\}_{i=1}^N, Q = \{q_j\}_{j=1}^M$ be point clouds in $\mathbb{R}^d$. The *Hausdorff distance* from P to Q is

$$d_H = \max\{\max_{p_i \in P} \min_{q_j \in Q} d(p_i, q_j), \max_{q_j \in Q} \min_{p_i \in P} d(p_i, q_j)\}.$$

That is, for each point in P (resp Q), we find the closest point in Q (resp P, and we take the largest such distance for all points in P (resp Q). The larger of these is the Hausdorff distance between the point clouds.

Theorem 1 (Stability theorem)**.** *Let P, Q be point clouds in $\mathbb{R}^d$, $\check{C}(P), \check{C}(Q)$ the Čech complexes, and $D(P)$ and $D(Q)$ the corresponding degree p persistence diagrams. Then*

$$d_B(D(P), D(Q)) \leq d_H(P, Q)$$

Thus, if we "wiggle" the points in a point cloud by ϵ, the diagrams will only change by at most ϵ.