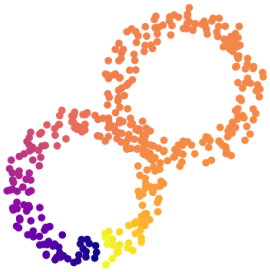
Topological Data Analysis Spring 2022: Assignment 1

Due: Tuesday April 5 (23:59)



Goal In this assignment you will implement *circular coordinates* and apply your implementation to a range of data sets.

Evaluation Each problem below has an associated weight and the sum of the weights add up to 100%. It suffices to answer the questions item by item: it is not necessary to write a "scientific report" where you include background material and discussions beyond what is explicitly asked for. You are free to produce the plots in whatever way you like and the efficiency of the implementations will not be scored. The circular coordinates must however be implemented using Ripser.py which is available at https://ripser.scikit-tda.org/en/latest/. **You do not need to submit code unless explicitly told to do so.**

The Assignment

In this assignment we shall see how persistent (co-)homology can be applied to parametrize a data set with respect to a "1-dimensional hole". The above figure shows an example of such coordinization. We give a brief summary of the mathematics underlying the theory of circular coordinates sufficient for solving the exercises.

Cohomology in Dimension 1

Let X be a simplicial complex and let $X^{[0]}, X^{[1]}$ and $X^{[2]}$ denote the sets of vertices, edges and triangles, respectively. We shall assume that the vertices $v_i \in X^{[0]}$ have a total order, and an edge connecting two vertices $v_i < v_j$ will be written v_iv_j. Similarly we denote a triangle by $v_iv_jv_k$ where $v_i < v_j < v_k$. Let A denote $\mathbb{Z}_p$ for some prime p, the integers $\mathbb{Z}$, or the real numbers $\mathbb{R}$. Define the following sets which are modules over A (i.e. vector spaces for $A \in \{\mathbb{Z}_p, \mathbb{R}\}$ and a finitely generated abelian group for $A = \mathbb{Z}$),

$$C^0(X; A) := \{\text{functions } f\colon X^{[0]} \to A\}$$
$$C^1(X; A) := \{\text{functions } \alpha\colon X^{[1]} \to A\}$$
$$C^2(X; A) := \{\text{functions } \Lambda\colon X^{[2]} \to A\}.$$

Define *co-boundary maps* $\partial^0 \colon C^0(X;A) \to C^1(X;A)$ and $\partial^1 \colon C^1(X;A) \to C^2(X;A)$ by

$$\partial^0(f)(v_1v_2) = f(v_2) - f(v_1),$$
$$\partial^1(\alpha)(v_1v_2v_3) = \alpha(v_2v_3) - \alpha(v_1v_3) + \alpha(v_1v_2).$$

If $\partial^1(\alpha) = 0$, then we say that α is a *cocycle*, and if $\alpha = \partial^0(f)$, then we say that α is a *coboundary*. It is straightforward to verify that $\partial^1 \circ \partial^0 = 0$ and thus $\operatorname{Im}\partial^0 \subseteq \ker\partial^1$. We define the *1-st cohomology group* to be the quotient module

$$H^1(X;A) := \frac{\ker \partial^1 \colon C^1(X;A) \to C^2(X;A)}{\operatorname{Im} \partial^0 \colon C^0(X;A) \to C^1(X;A)}. \tag{1}$$

We say that two cocycles α, β are *cohomologous* if $[\alpha] = [\beta]$ in $H^1(X;A)$.

Persistent Cohomology

Now we shall restrict our attention to the case $A = \mathbb{Z}_p$ for some prime p. In this case one can show that the vector space $H^1(X;\mathbb{Z}_p)$ is the vector space dual of $H_1(X;\mathbb{Z}_p)$ (this is also true for higher cohomology vector spaces) and therefore

$$\dim H^1(X;\mathbb{Z}_p) = \dim H_1(X;\mathbb{Z}_p).$$

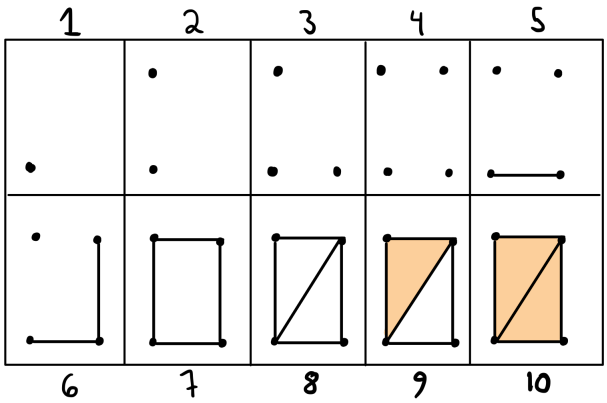
If $i \colon X \subseteq X'$ denotes the inclusion of a simplicial complex, then we get an induced morphism $i^* \colon H^1(X';\mathbb{Z}_p) \to H^1(X;\mathbb{Z}_p)$ defined by $i^*([\alpha]) = [\alpha \circ i]$. Note the direction of the induced morphism is reversed. In particular, for a filtration

$$X_0 \subseteq X_1 \subseteq \cdots \subseteq X_m = X,$$

we get an associated sequence of vector spaces and linear maps in cohomology:

$$H^1(X_0;\mathbb{Z}_p) \leftarrow H^1(X_1;\mathbb{Z}_p) \leftarrow \cdots \leftarrow H^1(X_m;\mathbb{Z}_p).$$

Just as in the case of homology, and as we saw in the lecture, such a collection of vector spaces and linear maps has a *barcode*. And since we are taking homology and cohomology with coefficients in a field, there is a one-to-one correspondence between the barcode in homology and the barcode in cohomology. Indeed, since $H^1(X_i;\mathbb{Z}_p) \cong \operatorname{Hom}(H_1(X_i;\mathbb{Z}_p),\mathbb{Z}_p)$ (the dual vector space), the diagram of Eq. (1) can be obtained (up to isomorphism) by first fixing a basis for every vector space in homology, and then transposing all the matrices representing the linear maps in homology. This equivalence between barcodes in homology and cohomology is particularly useful in topological data analysis as cohomology computations are typically much faster; state-of-the art software such as Ripser computes the barcode in cohomology rather than homology. Cohomology will also become essential when dealing with circular coordinates later in the assignment. In the lecture notes we computed the barcode and persistence diagram of the following filtration over $\mathbb{Z}_2$:



Its homology barcode in dimension 1 consists of two intervals: $[7, 10)$ and $[8, 9)$ and those bars have representative cycles $[v_1v_2 + v_1v_3 + v_2v_4 + v_3v_4]$ and $[v_1v_2 + v_2v_4 + v_1v_4]$, respectively. Here we have labeled the vertices by their order of appearance. In cohomology we are moving in the opposite order and thus there will be a feature which is born at 9 which vanishes as we enter 6, i.e. an interval $(6, 9]$ in the barcode, and similarly there will be an interval $(7, 8]$. However, it is customary to identify an interval $(a, b]$ in cohomology with its corresponding interval $[a+1, b+1)$ in homology. **Software such as Ripser will output the barcode/persistence diagram of the homology barcode**.

The representative cocycles on the other hand may look very different from the representative cycles we compute in homology. As an example, the function $\alpha \in C^1(X_9; \mathbb{Z}_2)$ defined on edges by $\alpha(v_1v_3) = 1$, and $\alpha(e) = 0$ for all other edges, satisfies $0 \neq [\alpha] \in H_1(X_9; \mathbb{Z}_2)$. Another possible generator would be the function which maps $v_3v_4 \mapsto 1$ and all other edges to 0. The function which maps $v_1v_2 \mapsto 1$ and all edges to 0 is not in the kernel of ∂^1 (why?). Furthermore, $\beta \in C^1(X_9; \mathbb{Z}_2)$ defined on edges by $\beta(v_1v_3) = \beta(v_3v_4) = 1$ and $\beta(e) = 0$ for all other edges is trivial in cohomology because $\beta = \partial^0(g)$ where $g(v_3) = 1$ and trivial on the other vertices. Also note that that the image of $[\alpha] \in H^1(X_9; \mathbb{Z}_2)$ in $H^1(X_6; \mathbb{Z}_2)$ vanishes as $\alpha = \partial^0(g') \in C^1(X_6; \mathbb{Z}_2)$ where $g'(v_1) = 1$ and $g'(v) = 0$ for all other vertices v.

Warm-up (5%)

In this section we shall visualize a representative cocycle for a point in the persistence diagram.

- Go to
  `https://ripser.scikit-tda.org/en/latest/notebooks/Representative%20Cocycles.html`
  and make sure you understand the example.
- Copy the code to your computer.
- Use the following code to sample 20 points from the circle.

  ```
  N=20
  t = np.linspace(0, 2*3.1415,N , endpoint=False)
  x = np.transpose([np.cos(t), np.sin(t)]) + 0.4*np.random.random((N,2))
  ```

  Project the representative cocycle of the maximum persistent point (i.e. longest bar) in dimension 1 onto the edges present right before the feature vanishes, and onto the edges present right after its birth (exactly as done on the webpage). Provide plots showing the projections in both cases together with the persistence diagram (in dimensions 0 and 1).

Circular Coordinates

There is a natural map $H^1(X; \mathbb{Z}) \to H^1(X; \mathbb{Z}_p)$ given by $[\alpha] \mapsto [\pi \circ \alpha]$ where $\pi \colon \mathbb{Z} \to \mathbb{Z}_p$ is the projection map which sends an integer to its congruence class. This map need not be an epimorphism but for a "random" prime p it will be. Indeed, given $[\alpha] \in H^1(X; \mathbb{Z}_p)$ we define a function

$$\hat{\alpha} \colon X^{[1]} \to \mathbb{Z}$$

by

$$\hat{\alpha}(\sigma) = \begin{cases} \alpha(\sigma) & \text{if } \alpha(\sigma) \leq \frac{p-1}{2} \\ \alpha(\sigma) - p & \text{if } \alpha(\sigma) > \frac{p-1}{2}. \end{cases} \tag{2}$$

If $\hat{\alpha}$ vanishes under the action of the boundary operator, then $[\hat{\alpha}]$ defines an element in $H^1(X;\mathbb{Z})$ which maps onto $[\alpha] \in H^1(X;\mathbb{Z}_p)$ under the above mapping. While that is not always true, it turns out to *always* be the case in practice for a large enough prime (e.g. $p = 41$).

The importance of integer coefficients stems from the fact every $[\alpha] \in H^1(X;\mathbb{Z})$ defines a homotopy class of continuous maps $|X| \to S^1$ in the following way: define $\theta\colon |X| \to S^1 = \mathbb{R}/\mathbb{Z}$ by mapping every vertex $v \in |X|$ to 0, and every edge e in $|X|$ around the entire circle with winding number $\alpha(e)$. The map is then by extended linearly to higher-order simplices. One can show that the resulting map is continuous and that cohomologous 1-cocycles define homotopic maps.

Such a map is not very "smooth" and we would like to "smooth out" the mapping while remaining in the same homotopy class. For any $f\colon X^{[0]} \to \mathbb{R}$ we can distribute the image of the vertices of X along the circle by mapping every vertex to $f(v) \pmod{\mathbb{Z}}$, and by "stretching" the edges accordingly. I.e. the edge v_iv_j is mapped to to the line of (signed) length $\alpha(v_iv_j) + f(v_j) - f(v_i)$. The "smoothing process" chooses an f such that the square sum of the edge lengths becomes minimal. The fact that this is always possible can be justified by studying cohomology with real coefficients.

First observe that the inclusion $\iota\colon \mathbb{Z} \hookrightarrow \mathbb{R}$ induces a homomorphism $H^1(X;\mathbb{Z}) \to H^1(X;\mathbb{R})$ defined by $[\alpha] \to [\iota \circ \alpha]$. For every $\beta \in H^1(X;\mathbb{R})$, define

$$||\beta||^2 = \sum_{v_iv_j \in X^{[1]}} \beta(v_iv_j)^2. \tag{3}$$

Importantly, one can show that there for every $\alpha \in H^1(X;\mathbb{Z})$ exists a *unique* $\beta \in H^1(X;\mathbb{R})$, cohomologous to $\iota \circ \alpha$, such that Eq. (3) is minimal. Since β is cohomologous to $\iota \circ \alpha$, we must have that $\beta = \iota \circ \alpha + \partial^0(f)$ for some function $f\colon X^{[0]} \to \mathbb{R}$. We define the *circular coordinates associated to* α to be the function $\theta\colon X^{[0]} \to \mathbb{R}/S^1$ given by $\theta(v) = f(v) \pmod{\mathbb{Z}}$.

Computation Consider the simplicial complex X given by the boundary of a triangle, with vertices $\{v_0, v_1, v_2\}$ ordered by $v_0 < v_1 < v_2$. Let $\alpha\colon X^{[1]} \to \mathbb{Z}$ be given by $\alpha(v_0v_1) = 1$ and $\alpha(v_0v_2) = \alpha(v_1v_2) = 0$. We shall now find the circular coordinates associated to α. By representing the coboundary matrix $\partial^0\colon C^0(X;\mathbb{R}) \to C^1(X;\mathbb{R})$ in the bases given by the simplices we get

$$\partial^0 = M = \begin{array}{c} \\ v_0v_1 \\ v_0v_2 \\ v_1v_2 \end{array} \begin{array}{c} \begin{array}{ccc} v_0 & v_1 & v_2 \end{array} \\ \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & 1 \end{pmatrix} \end{array}.$$

Denoting the function values $f(v_i)$ by y_i, and letting $b = (1,0,0)^T$ be α in the basis given by the 1-simplices, we are seeking a vector $y^T = (y_0, y_1, y_2)^T$ such that

$$||\iota \circ \alpha + \partial^0(f)||^2 = ||b + My||_2^2$$

is minimal. But this is simply a least squares optimization problem! It follows that $y = M^\dagger(-b)$ where $M^\dagger$ is the Moore-Penrose pseudoinverse of M. Recall that $M^\dagger$ can be computed as $V \cdot \Sigma^{-1} \cdot U^T$ where $M = U \cdot \Sigma \cdot V^T$ is the singular value decomposition of M.

In the above example one finds that

$$M^\dagger = \begin{bmatrix} -1/3 & -1/3 & 0 \\ 1/3 & 0 & -1/3 \\ 0 & 1/3 & 1/3 \end{bmatrix},$$

and thus $y = M^\dagger(-b) = [1/3, -1/3, 0]^T$. By working modulo $\mathbb{Z}$ this gives $f(v_0) = 1/3, f(v_1) = 2/3$ and $f(v_2) = 0$. Or, if you prefer to identify the circle with $[0, 360)$, we get the coordinates $[120, 240, 0]$ - i.e. three equidistant points on the circle.

This approach generalizes readily to any 1-cocycle and any simplicial complex X.

Algorithm We summarize how to compute circular coordinates.

1. Compute persistent cohomology with coefficients in $\mathbb{Z}_p$ for a large p (e.g. $p = 41$).
2. Choose a point (x, y) in the persistence diagram and extract a representative cocycle α for that feature.
3. Fix a scale $r \in [x, y)$ and project α onto the edges v_iv_j with $d(v_i, v_j) \leq r$. Denote the projection by $\alpha_r \in H^1(X_r; \mathbb{Z}_p)$.
4. Use Eq. (2) to lift α_r to a cocycle $\widehat{\alpha_r} \in H^1(X_r; \mathbb{Z})$.
5. Represent the coboundary operator $\partial^0 \colon C^0(X_r; \mathbb{R}) \to C^1(X_r; \mathbb{R})$ in the basis given by the simplices (this is just the transpose of the boundary matrix). Denote the resulting matrix by M and compute its Moore-Penrose pseudoniverse $M^\dagger$.
6. Write $\widehat{\alpha_r}$ as a vector in the basis given by the 1-simplices (in the same order as used when constructing M). Compute $y = M^\dagger(-b)$.
7. The circular coordinate of the vertex v_i is $\theta_i \in [0, 1)$ where $\theta_i = y_i \mod \mathbb{Z}$.

Implementation (40%)

- Implement a function in python that takes as input a tuple (D, α, r, p) where
  - D is an $n \times n$-matrix of pairwise distances between n data points,
  - α is a representative cocycle for a point (x, y) in the persistence diagram of D,
  - $r \in [x, y)$ is the threshold at which α is lifted to an integer cocycle,
  - p is the prime used to compute the persistence diagram,

  and returns the circular coordinates of the n data points in a list.
- Submit your code!

First Example (5%)

- Use the following code to generate data

```
N=100
t = np.linspace(0, 2*3.1415,N , endpoint=False)
x = np.transpose([np.cos(t), np.sin(t)]) + 0.2*np.random.random((N,2))
```

- Compute the circular coordinates associated to the representative 1-cocycle of the most persistent point in the persistence diagram.
- Produce a scatter plot (like the one on Page 1) where each point is color-coded by its circular coordinate.
- Submit the code!

The Flat Torus (15%)

- Write a function that takes as input a set of points in $[0,1] \times [0,1]$ and outputs the distance matrix D of the points considered as a subset of the flat torus. That is, $(u, 0)$ is identified with $(u, 1)$ and $(0, v)$ is identified with $(1, v)$. Do submit the code!

- Sample 225 points from the flat torus using the following code

```
n=15
t = np.linspace(0,1,n)
x,y = np.meshgrid(t,t)
x = np.reshape(x, (n*n,1))
y = np.reshape(y, (n*n,1))
X = np.hstack([x, y]) + 0.05*np.random.random((n*n,2))
```

  Produce scatter plots as in the previous exercise for the *two* most persistent points. Which threshold r did you choose, and why?

Image Data (20%)

The Columbia University Image Library (COIL-20) is a collection of 448×416-pixel gray scale images from 20 objects, each of which is photographed at 72 different rotation angles. The database has two versions: a processed version, where the images have been cropped to show only the rotated object, and an unprocessed version with the 72 raw images from 5 objects.

- Downlad the unprocessed version from
`https://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php`

- Considering each image as a vector of pixel intensities in $\mathbb{R}^{448\times416}$ yields a data set of 72 points for each object. For an object of your choice, compute the associated distance matrix (using the Euclidean distance) and compute the circular coordinates associated to the 1-cocycle representing the most persistent point. Produce a plot showing the images together with their inferred circular coordinates.

- Now consider the data from all images simultaneously, i.e. 360 data points in $\mathbb{R}^{448\times416}$. Compute the associated distance matrix and persistence diagram, and order the pairs (x, y) in the persistence diagram by the quantity $\frac{y-x}{x}$. Compute the circular coordinates associated to the representative cocycles of the five greatest pairs. Use *principal component analysis* to project the images from $\mathbb{R}^{448\times416}$ down onto $\mathbb{R}^2$. For each of the five circular coordinates, produce a scatter plot where each point is color-coded by its circular coordinate.

Surprise Me (15%)

- Inspired by the previous data sets, construct a data set of your own and plot the associated circular coordinates in a reasonable way. Explain carefully how the data was sampled. Points will be awarded based on creativity (i.e. a uniform sampling of the circle is unlikely to generate many points).