Optimal cycles and applications in machine learning

Carlos Henrique Venturi Ronchi, Marcio Fuzeto Gameiro

Institute of Mathematical and Computer Sciences - University of São Paulo

Introduction

In recent years data is being produced at an unprecedent rate. Persistent homology can be used to help in the task of visualizing and understanding the shape of data, such as detecting path connected components, holes and cavities of our data. Persistent homology has undergone significant progress in recent years. Lately many algorithms have been developed to increase performance and to better understand the shape of data. One way to achieve the former is to consider the representatives (cycles) of homology groups and persistent homology filtration and try to optimize them using linear programming [2].

Persistent homology

Definition: *A persistence module* $\mathbb{V}$ *over the real numbers* $\mathbb{R}$ *is defined to be an indexed family of vector spaces* $(V_t | t \in \mathbb{R})$ *and a doubly-indexed family of linear maps* $(v_t^s : V_s \to V_t | s \leq t)$ *which satisfy the composition law* $(v_t^s \circ v_s^r = v_t^r)$ *whenever* $r \leq s \leq t$, *and where* v_t^t *is the identity map on* V_t.

Let K be a simplicial complex and $K_0 \subset K_1 \subset \cdots \subset K_n = K$ a filtration of K, where each K_t is a subcomplex.

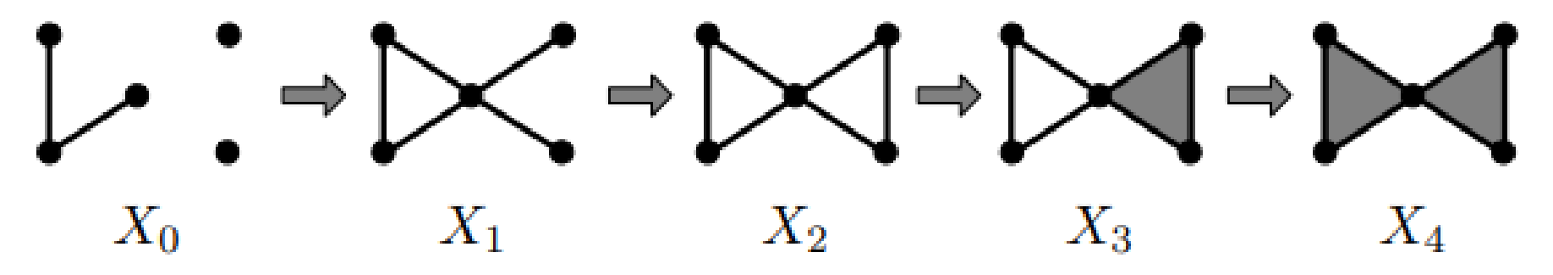

Figure 1 Filtration of a simplicial complex

It follows that the vector spaces $H(X_t)$ are finite-dimensional and for a finite set $a_1 < \cdots < a_m$ of "critical values" for the simplicial complex K all the information of the persistence module $(H(X_t), v_t^s = H(i_t^s))$ is encoded in the following finite diagram:

$$H(X_{a_1}) \to \cdots \to H(X_{a_m})$$

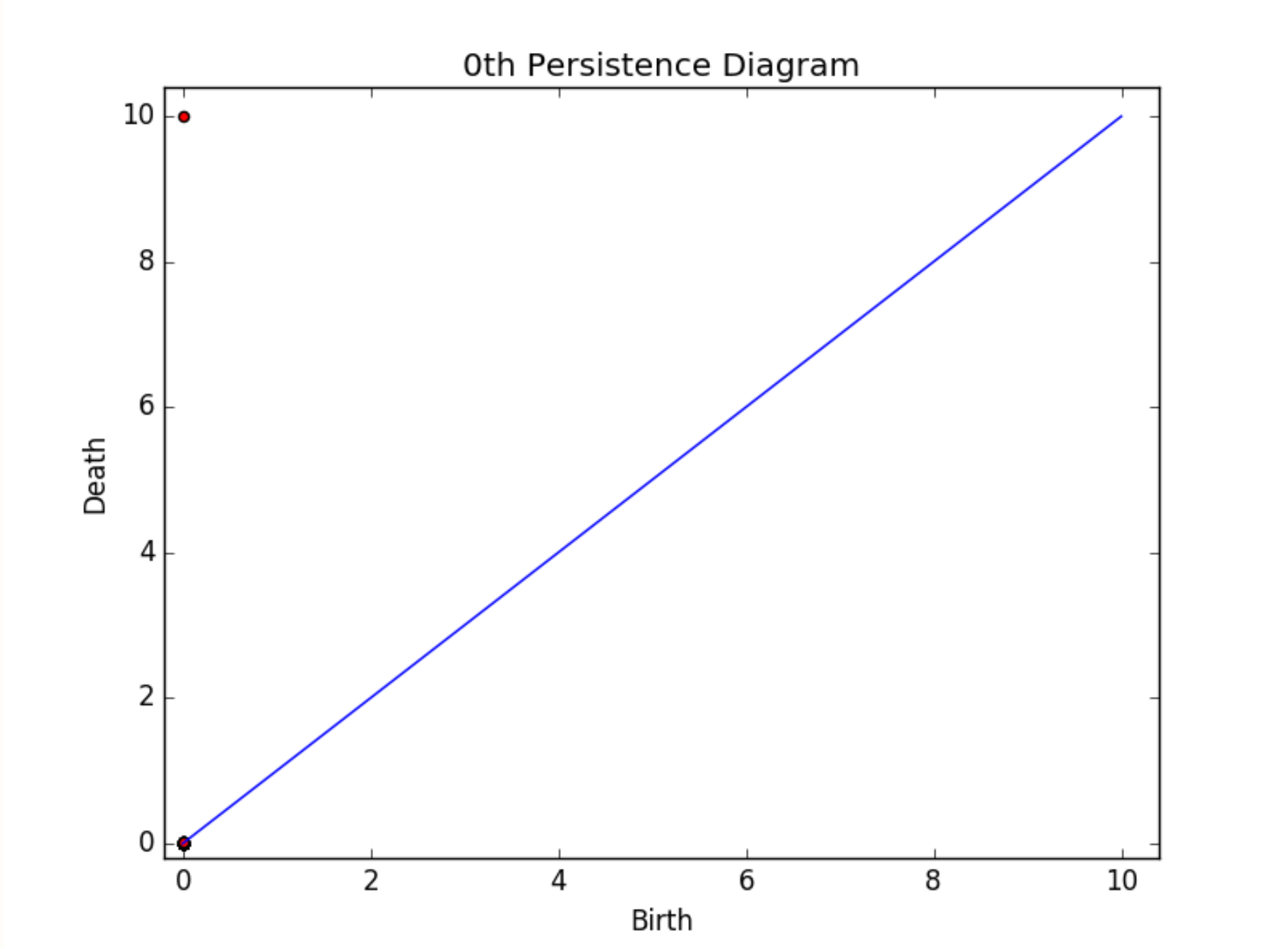

Figure 2 0th Persistence diagram of torus

Definition: *The above description is the persistence diagram, or barcode of the simplicial complex* K.

We may generalize the idea of persistence diagram for decomposable persistence modules. First we need the following theorem.

Theorem: *Let* $\mathbb{V}$ *be a persistence module over* $T \subset \mathbb{R}$. *Then* $\mathbb{V}$ *can be decomposed as a direct sum of interval modules in either of the following situations:*

1. T *is a finite set;*
2. *each* V_t *is finite-dimensional.*

On the other hand, there exists a persistence module over $\mathbb{Z}$ *(indeed, over the nonpositive integers) which does not admit an interval decomposition.*

Given a decomposable persistence module $\mathbb{V}$ indexed over $\mathbb{R}$,

$$\mathbb{V} \cong \oplus_{l \in L} \mathbf{k}(p_l^*, q_l^*)$$

then we define the persistence diagram to be the multiset

$$\operatorname{dgm}(\mathbb{V}) = \{(p_l, q_l) | l \in L\} - \Delta$$

where Δ is the diagonal in the plane.

Loosely speaking the persistence diagram tells us the topological properties which are *born* at the index p_l and *die* at the index q_l.

For further details regarding persistence modules, check [1].

Optimal cycles for persistent homology

When calculating the persistent homology we generate a set of representatives from the homology groups, but they are not as optimal as they could be. Given a cycle $z \in Z_q(X)$ we can consider the following problem:

$$\begin{array}{ll} \text{minimize} & \|x\|_1 \\ \text{subject to} & \begin{cases} x - \partial_{q+1} y = z \\ x, y \text{ integral} \end{cases} \end{array}$$

where the 1-norm is defined as $\|\sum \alpha_i \sigma_i\|_1 = \sum |\alpha_i|$. The solution $\tilde{z}$ to the above optimization problem is called an optimal cycle homologous to z.

Let $K_0 \subset K_1 \subset \cdots \subset K_n = K$ be a filtration where only one simplex is added at each index, we can define the set of cycles $\{g_1, \ldots, g_n\} \subset H_q(X_k)$ of dimension q that are not a boundary of a $q+1$-chain in X_k. One can show that the set $\{g_1, \ldots, g_n\}$ forms a basis for $H_q(X_k)$. We can optimize the cycles after they are born using the following procedure.

```
Algorithm 1: Optimize cycles
Given $z_j = g_j$, find an optimal solution to $\tilde{z}_j$ to
    minimize      $\|x\|_1$
    subject to    $\{ x + By + \sum_{i \in \mathscr{L}_q(j), i<j} a_i \tilde{z}_i = z_j$
```

where $\mathscr{L}_q(j)$ is the set of indices of those cycles g_i of dimension q that are not part of the boundary in X_j. To calculate each g_j we initialize the usual persistence algorithm with $g_j = \sigma_j$ and procede as usual. It then suffices to show that this new cycles $\tilde{z}_j$ form a basis for $H_q(X_k)$.

Theorem: *Given the output of the above algorithm,* $\{[\tilde{z}_i] | i \in \mathscr{L}_q(k)\}$ *forms a basis for* $H_q(X_k)$

Sketch: first note that from the algorithm $[\tilde{z}_i] = [g_i] + \sum_{h \in \mathscr{L}_q(i), h<i} a_h \tilde{z}_h$ and for those cycles $[\tilde{z}_h']$ that die before entering X_k we have $[g_h'] = 0$, where $h' < h$ for each h in the above summand. We then get

$$[\tilde{z}_i] = [g_i] + \sum_{h \in \mathscr{L}_q(i), h<i} c_h [g_h]$$

and since this is an invertible transformation we can simply change $[g_i]$ for $[\tilde{z}_i]$.

We now show an example. Consider the following figure.

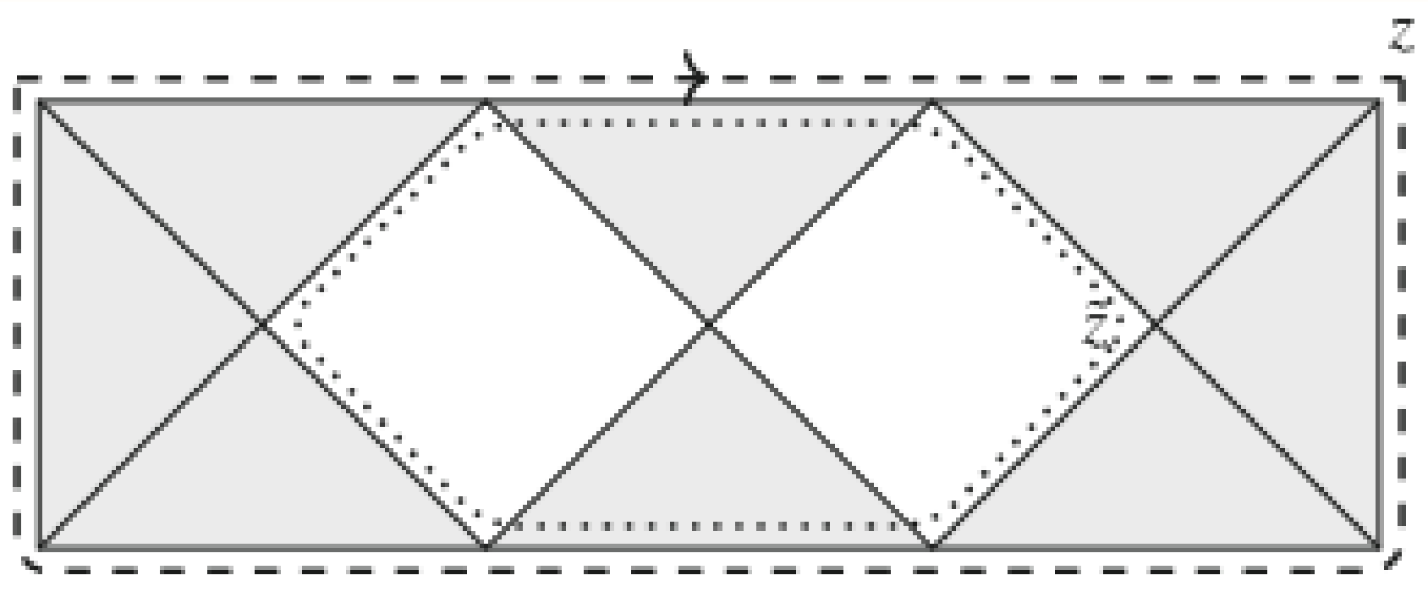

Figure 3 Homologous cycles z and $\tilde{z}$. (Obtained from [2])

When applying the algorithm for a single cycle z gets optimized to $\tilde{z}$. Although we want to detect both holes. Let z_1 be the outer dashed cycle and z_2 the left cycle that goes around the left hole. When we apply the algorithm for multiply cycles we get $\tilde{z}_1$ to be the right cycle. The idea is when we add the simplex whose boundary is z_2 we fill the hole and obtain $\tilde{z}_1$.

Optimal cycles and machine learning

A classifier of tourist attractions in Curitiba, Brazil was obtained using a convolution neural network, Alexnet architecture [3]. The proposed classifier achieved an accuracy of 70%. In order to improve the accuracy, we tracked the optimal cycles in some images from specific classes, so we could train other classifiers with fewer classes.

Figure 4 Touristic attraction images used to calculate the optimal cycles.

For each image we transformed it in a point cloud, where each pixel corresponded to a point in the plane. Afterwards we calculated their respective persistence diagram using the alpha shape filtration and the software *OptiPersLP* based on the article [2]. Analyzing each diagram we chose the point with largest lifespan and looked up their respective optimized cycle in the plane.

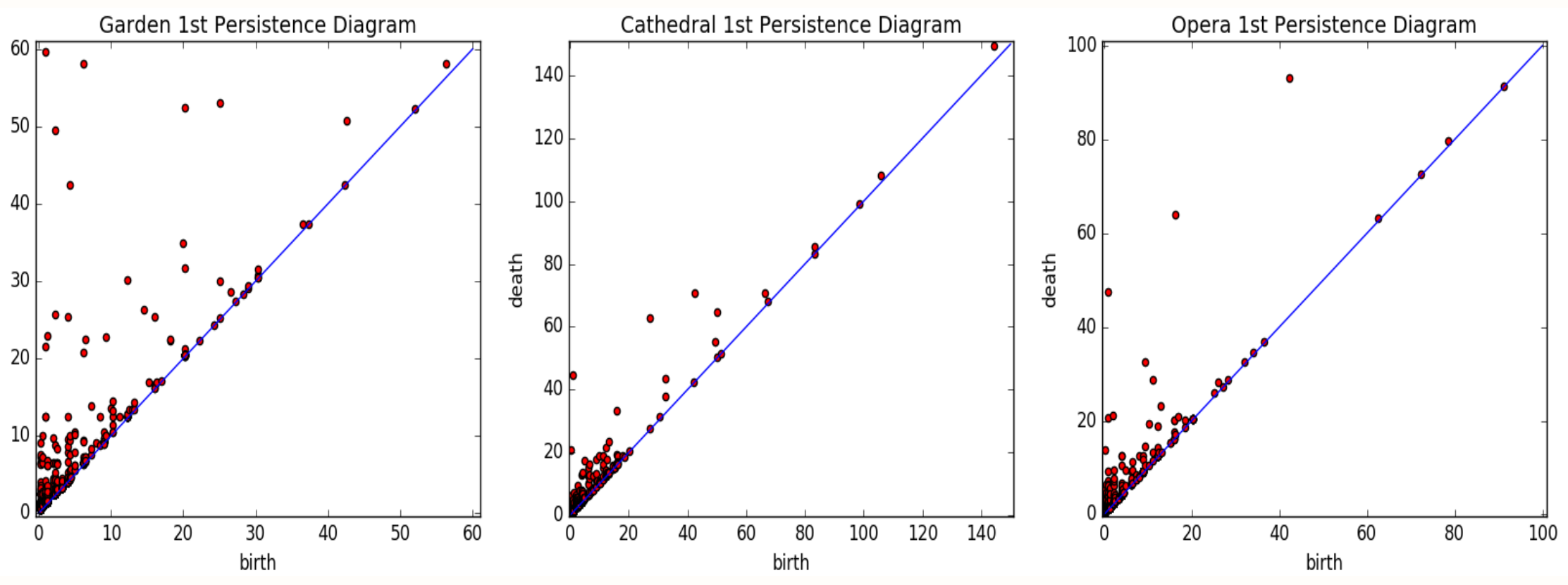

Figure 5 The correspondent persistence diagram for each point cloud generated from the images.

We then extracted squared patches that did not meet the holes found with the persistent homology, since they tend to be more meaningful as they do not contain significant holes and after that trained three new classifiers. Ensembling the new classifiers with the older one we improved the accuracy in 5%.

Table 1 Accuracy of the classifiers using the described method. They are only two classes for each classifier.

Classes	Classifier	Accuracy
Garden and Cathedral	RBF SVM	68.29%
Garden and Opera	RBF SVM	78.04%
Cathedral and Opera	RBF SVM	73.17%

References

[1] Frédéric Chazal et al. *The Structure and Stability of Persistence Modules.* Springer International Publishing, 2016. DOI: 10.1007/978-3-319-42545-0. URL: https://doi.org/10.1007/978-3-319-42545-0.

[2] Emerson G. Escolar and Yasuaki Hiraoka. "Optimal Cycles for Persistent Homology Via Linear Programming". In: *Optimization in the Real World.* Springer Japan, Sept. 2015, pp. 79–96. DOI: 10.1007/978-4-431-55420-2_5. URL: https://doi.org/10.1007/978-4-431-55420-2_5.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems 25.* Ed. by F. Pereira et al. Curran Associates, Inc., 2012, pp. 1097–1105. URL: https://goo.gl/yfJonT.