# Enhanced Persistence in Topological Data Analysis

Shirley Toribio

Advisor: Prof. Chandler

September 1, 2023

## 1 Introduction

Topological data analysis (TDA) is a rapidly growing field at the intersection of applied algebraic topology, computational geometry, statistics, and data science. Persistent homology is a particularly popular tool within TDA that's meant to track the persistence of features within an object. Persistence diagrams are an important tool for representing different objects and their resulting filtrations (to be defined below), and due to their importance, bottleneck distance, a measure of the difference between two persistence diagrams, has become an important tool within TDA. In order to better improve the information measured by bottleneck distance, an enhanced form of bottleneck distance that considers multiple dimensions of structures in a disciplined way is considered. The second section of this paper reviews important concepts for understanding persistent homology and bottleneck distance; the third section of this paper discusses the motivation behind developing an enhanced persistence metric and its theoretical underpinnings; the fourth section reviews the methods and tools used to analyze the effectiveness of an enhanced bottleneck metric; the fifth section reviews the analysis done on this enhanced metric; the sixth section concludes this paper with the limits of this research and future possibilities involving this enhanced metric.

## 2 Conceptual Overview

TDA involves analyzing the topological and geometric properties underlying the structure of data and using them to extract qualitative and quantitative information about said data. Topological properties are shared among shapes or structures that are homeomorphic to each other, meaning they can be transformed into one another via deformation, with a few exceptions. One classic example of this involves a donut-like shape and a mug-like shape. Assuming both objects are malleable, one could be deformed into the other and vice versa. TDA allows for the identification and analysis of an underlying shape in a sample of data that might be useful for drawing conclusions about said data.

Simplicial complexes are an important tool in TDA. A simplicial complex is a set composed of points, line segments, and triangles (and potentially higher dimensional elements that are not used in the present work). The simplest simplex type, a point, is also referred to as a 0-dimensional simplex. This is followed by a 1-dimensional simplex, or line, and a 2-dimensional simplex, which is a triangle. A 3-dimensional simplex corresponds to a tetrahedron, and so on to an n-dimensional simplex. Every n-dimensional simplex is bordered by $n + 1$ $(n - 1)$-dimensional simplices, which are called its faces. For instance, a 2-dimensional simplex (or triangle) is bordered by 3 1-dimensional simplices just as a line, or a 1-simplex, is bordered by 2 points, or 0-dimensional simplices.

Although simplicial complexes are composed of simplices, arbitrary collections of simplicies are not necessarily a simplicial complex. For a set $\mathcal{K}$ to be a simplicial complex, we require the following conditions:
1. Every face of a simplex from $\mathcal{K}$ is also in $\mathcal{K}$.
2. The non-empty intersection of any two simplices $\sigma_1, \sigma_2 \in \mathcal{K}$ is a face of both $\sigma_1$ and $\sigma_2$.

Persistent homology is an important tool in TDA that involves the usage of simplicial complexes. It tracks both the changes in and persistence of topological properties in an object across different scales. This is done through a filtration, which is a family of simplicial complexes nested in one another. It can also be thought of as a large simplicial complex with nested subcomplexes within it, each corresponding to the same object at specific scales. It is important to note that each subcomplex has an assigned value within a filtration. The parameters of the scale can vary depending on the data being analyzed and choices made by the person performing the analysis.

Filtrations defined on a function can be separated into two different categories: sublevel filtrations and superlevel filtrations. Sublevel filtrations track the topological properties of an object as a parameter increases, with the subcomplex at a specific scale consisting of only those simplices with a value less than or equal to a certain value. Superlevel filtrations track the topological properties of an object as a parameter decreases and have subcomplexes that consist of only those simplices with a value greater than or equal to the value at a specific scale. Figure 1 shows a filtration for a function defined on a simplicial complex.

TDA focused on point clouds will often use a measure of distance between different points as its scaled parameter. The TDA described in this paper, however, is focused on analyzing functions instead of point clouds. For this reason, rather than the filtration being based on a set of points and their distances from

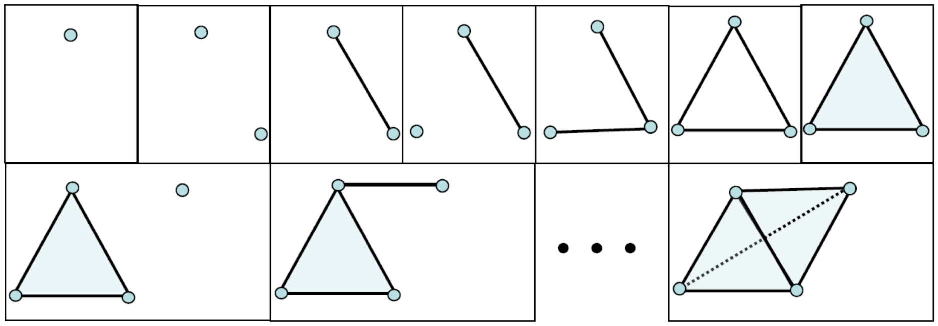

Figure 1: Filtration example. This could come about by a function defined on each 0-simplex in the full complex shown in the bottom right. Image via Chazal (2017).

each other, it is defined on a function.

Persistence diagrams are graphs used to visually portray the persistence of different topological features in a filtration. The $x$-coordinate of a point represents the value at which a specific topological feature was created, and the $y$-coordinate represents the value at which a specific topological feature ceased to exist. Points on the $x = y$ line thus correspond to topological features that never existed. See figure 2.

The components tracked by a persistence diagram depend on both the dimensionality of the object being analyzed and the focus of the person performing the analysis. They also tend to be sorted and labeled based on their dimensionality. 0-dimensional components are connected components, 1-dimensional structures are holes formed by 1 simplices, and 2-dimensional structures are voids, i.e. regions bounded by 2-dimensional simplices.

Persistence diagrams are thus a graphical summary of the homological persistence of a filtration. Given two persistence diagrams, $X$ and $Y$ (i.e. two sets of ordered pairs, i.e. birth and death times of topological features), bottleneck distance, $W(X, Y)$, is a measure of the distance between. In order to calculate the bottleneck distance, first a minimal bijective, or one to one, matching $\phi$ would need to be found between sets A and B based on the $L_\infty$ distance between points. The bottleneck distance would be the maximum $L_\infty$ distance between two points in this minimal bijective matching.

$$W_\infty(X, Y) := \inf_{\varphi : X \to Y} \sup_{x \in X} \|x - \varphi(x)\|_\infty$$

The $L_\infty$ distance is the maximum difference among the differences between $x$-coordinates and the difference between $y$-coordinates for two points.

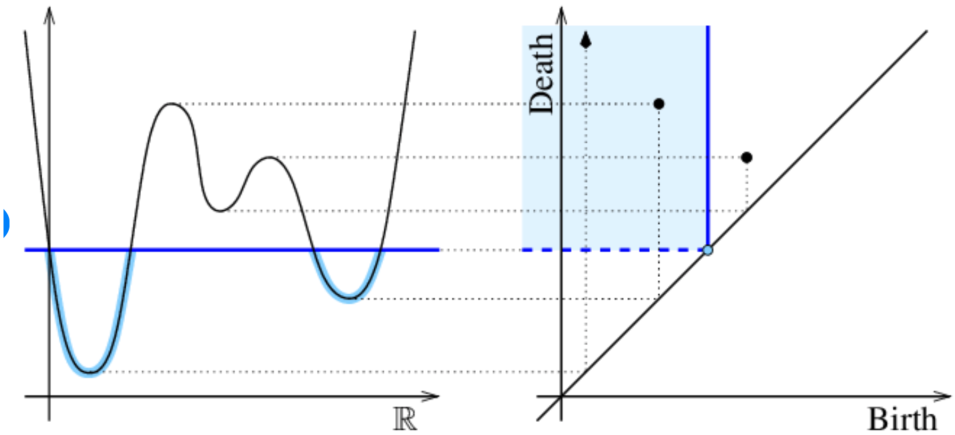

Figure 2: The persistence diagram coming from a sub-level set filtration on the function $f(x)$. Image via Edelsbrunner et al (2019).

## 3 Enhanced Persistence

Typically, in comparing a collection of data objects via their persistence diagrams, the scientist makes a decision of which dimensional structures to compare in calculating the bottleneck distance. In fact, the R function *bottleneck* in the TDA library requires an explicit choice of dimension. While certainly one can combine the bottleneck distance in multiple dimensions, information is lost in doing so. Holes found in the filtration exist within particular connected components. Consider the two objects “18” and “96”. Both consist of two connected components, and two holes. However they are topologically different, as “18” consists of both holes existing in the same connected component.

The idea of enhanced persistence is to connect connected components with their high dimensional homology. This is accomplished by considering the connected components coming from a superlevel set filtration, and then individually computing the persistence diagram with respect to one-dimensional homology for each of these connected components.

Consequently, we view a persistence diagram of 0-dimensional points as a diagram consisting of not just birth and death times $(b_i, d_i)$ but also its implicit homology: $(b_i, d_i, A_i)$ where $X$ is a persistence diagram for 1-dimensional holes.

We then adapt the bottleneck distance to this new high dimensional object as follows:

$$W_\infty(X, Y) := \inf_{\varphi : X \to Y} \big[ \sup_{x \in X} \|x - \varphi(x)\|_\infty + \gamma W_\infty(A, \phi(A)) \big] \tag{1}$$

# 4 Methods

The programming language R was used to perform calculations and algorithms, generate functions, and conduct persistent homology analysis. Two programming libraries besides base R were used: the TDA and igraph libraries. Concepts from graph theory were also used in performing calculations, so it is necessary to go over them for full understanding of the algorithm used to calculate enhanced bottleneck distance. A bipartite graph is a matching between two sets such that no two elements within the same set are matched to each other. For example, elements in set A match to elements in set B, but no two elements in set A are matched to each other and no two elements in set B are matched to each other. A maximal bipartite graph is a bipartite graph with the largest possible number of matchings.

In order to calculate enhanced bottleneck distance, an algorithm needed to be written in code first. The following algorithm for calculating bottleneck distance is detailed via Hellmer (2021):
1. Compute bipartite graph detailing all possible pairings between points in set A and points in set B. Set A consists of all off-diagonal points in persistence diagram A and the closest diagonal projections of all off-diagonal points in persistence diagram B. Set B consists of all off-diagonal points in persistence diagram B and the closest diagonal projections of all off-diagonal points in persistence diagram A.
2. Assign costs to each pairing. The cost of pairings between off-diagonal points from differing sets is their L-infinity distance, and the cost between off-diagonal points and their closest diagonal projections is also their L-infinity distance. Pairings between diagonal projections have a cost of 0 because these pairings could never represent the bottleneck distance.
3. Sort these costs in order of smallest to largest and perform a binary search for the bipartite match with the smallest cost sum among all possible bipartite matches that include all off-diagonal and diagonal points in their pairings. The binary search is performed by filtering out matchings that are greater than a certain cost from the maximum bipartite matching per each iteration.

Hierarchical clustering was used to assess the effectiveness of the enhanced bottleneck metric. This is an unsupervised algorithm that groups objects together using a distance matrix. Objects that are in the same group or subgroup are more closely associated with each other than objects that are in different groups.

# 5 Results

In order to assess the effectiveness of an enhanced bottleneck metric in measuring function distance, the best way to measure normal bottleneck distance

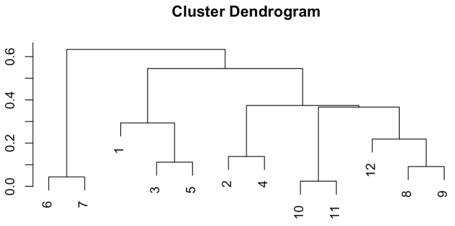

Figure 3: Example Dendrogram. The y-axis is a measure of the distance between different groups.

was first assessed. To do this, 20 functions were simulated, with functions 1-10 pertaining to functions of the same type and functions 11–20 pertaining to functions of another type.

Specifically, functions were generated as follows:

$$f(x,y) = \frac{1}{2}\sin(20y) + y^2 + \frac{1}{20}g(x,y;\mu_1) + \frac{1}{20}g(x,y;\mu_1)$$

where $g(x,y;\mu)$ is a bi-variate normal density centered at $\mu$ with covariance matrix $\Sigma = .04^2I$. $\mu_1 \sim N((.4,.4),.02^2I)$ for observations from class 1 and $\mu_1 \sim N((.4,.65),.02^2I)$ for observations from class 2. $\mu_2 = (0.67, 0.67)$ for both classes. Examples of these functions can be seen in figures 5 and 7.

Afterwards, 4 matrices of bottleneck distances between the different functions were formed. Each matrix corresponded to bottleneck distances assessed with a different type of filtration and dimensional component. Normal bottleneck distance was assessed using these four different parameters: 0-dimensional components and superlevel filtration, 1-dimensional components and superlevel filtration, 0-dimensional components and sublevel filtration, 1-dimensional components and superlevel filtration. For each set of bottleneck distances, functions were grouped together using hierarchical clustering. Only the dendrogram resulting from bottleneck distances assessed using 0-dimensional components and superlevel filtration successfully grouped the functions based on their type.

In order to calculate enhanced bottleneck distance, the cost of pairings between off-diagonal points and pairings between off-diagonal points and their diagonal projection became the sum of their $L-\infty$ distance and the normal bottleneck distance between the persistence diagrams of each point's corresponding connected component. For each connected component coming from the superlevel set filtration, the largest subcomplex for that component was found. See figure 4. Then, the persistence diagram for 1-dimensional features was computed for each of these sub-complexes. Bottleneck distance was then computed using equation (1) with $\gamma = .5$.

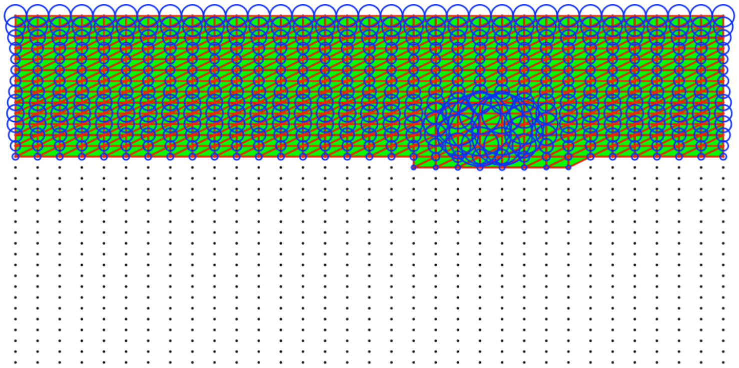

Figure 4: A subcomplex corresponding to the connected component generated from superlevel set filtration just before being absorbed into the connected component generated by the second mode as in figure 7. The persistence diagram corresponding to 1-dimensional structures based on the sublevel set filtration will exhibit one hole corresponding to the "bump."

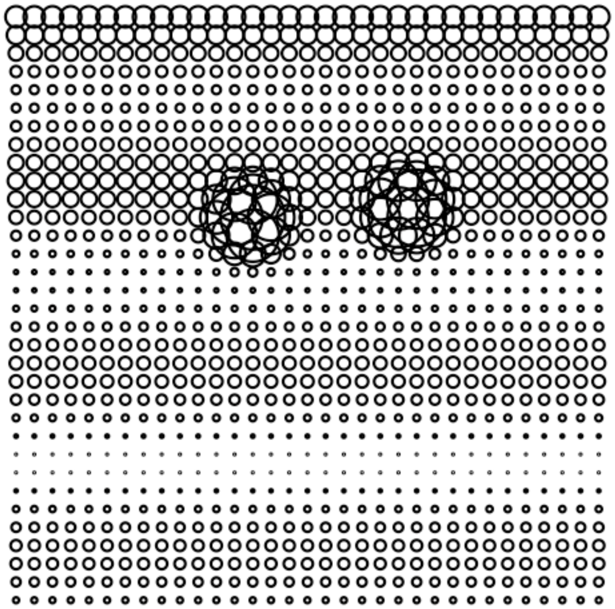

Figure 5: A function for which two normal "bumps" live on the same ridge.

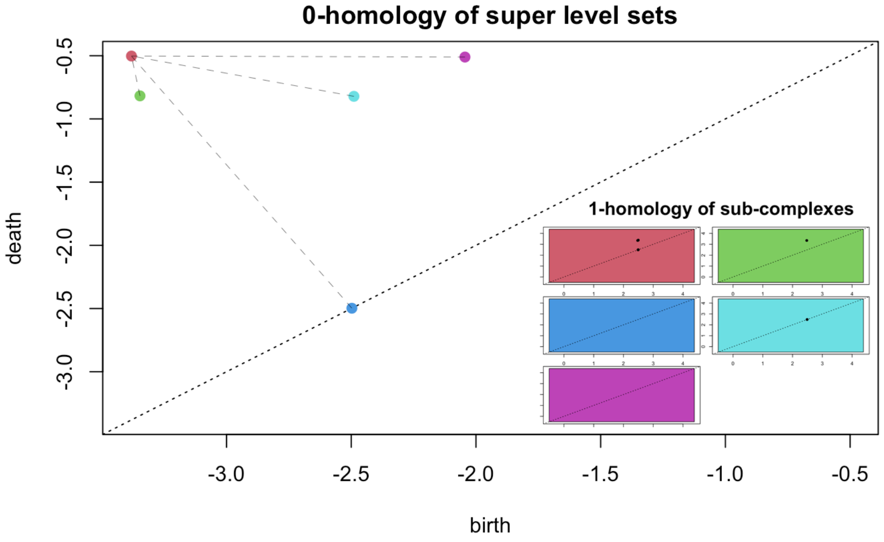

Figure 6: An enhanced persistence diagram where each colored point corresponding to a connected component is enhanced with the 1-dimensional persistence diagram of the same color. Lines correspond to which connected component each connected component merged into at “death” time. The underlying function is given in figure 5.

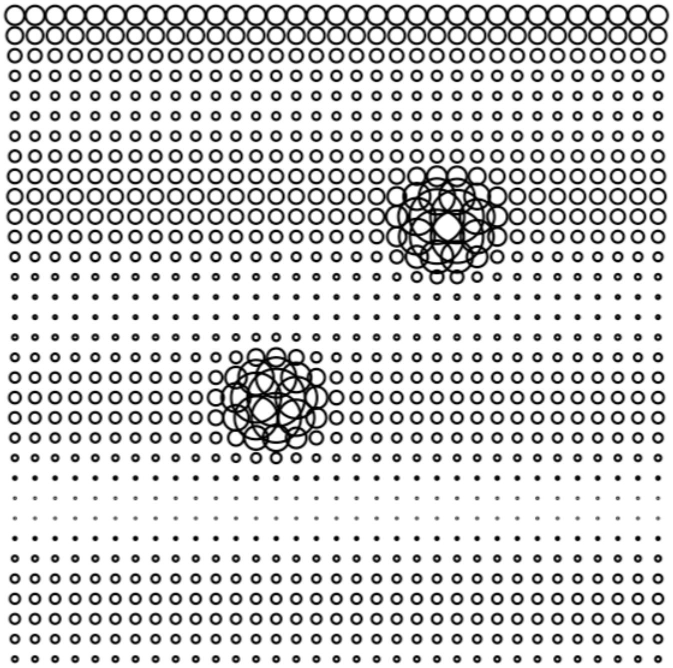

Figure 7: A function for which two normal “bumps” live on different ridges.

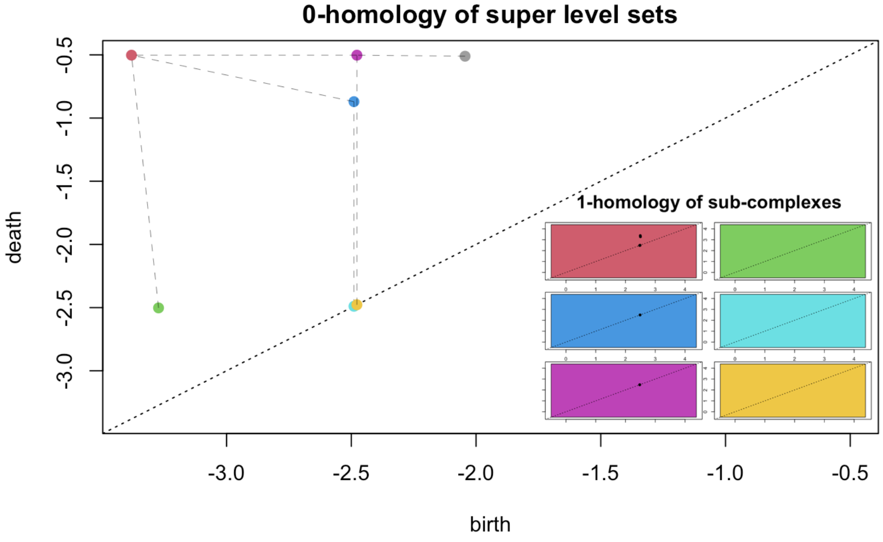

Figure 8: The enhanced persistence diagram for a function similar to figure 7

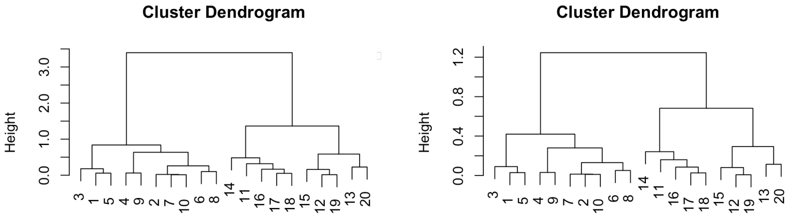

Figure 9: Cluster Dendrograms of Functions. The dendrogram on the left corresponds to bottleneck distances with enhanced persistence and the one on the right corresponds to normal bottleneck.

Although both types of bottleneck distance resulted in the 20 functions being accurately sorted based on function type, the cluster dendrogram resulting from the bottleneck distance using enhanced persistence showed a greater difference between the two clusters than the dendrogram resulting from normal bottleneck distance. This can be seen by the y-axis on both dendrograms. Enhanced persistence thus seems to improve the ability of bottleneck distance to distinguish between different types of functions. More analysis would need to be run to confirm this, however.

# 6 Conclusion

Bottleneck distance calculated with enhanced persistence presents a possible avenue for both increasing the amount of information provided by it and improving it as a metric. The analysis conducted in this paper serves as more of a preliminary analysis of its potential. More research is needed to see if bottleneck distance would benefit from the addition of enhanced persistence. Conducting more simulations with a larger variety of functions could be a potential research possibility. Assessing enhanced persistence alongside distance functions could also be useful in seeing if its use could be extended to the analysis of point clouds.

# 7 Works Cited

Chazal, Frederic. (2017). Persistent homology for TDA.

Edelsbrunner, Herbert, Virk, Ziga and Wagner, Hubert. (2019). Topological Data Analysis in Information Space.

Hellmer, Niklas. (2021). Computing the Bottleneck Distance.

## Acknowledgement

I would like to extend my gratitude to the Kenneth Cooke Memorial Endowment Fund for supporting my research this summer and my professors in the Department of Mathematics and Statistics for their continued support and guidance.