Sampling Bounds for Topological Descriptors

Brittany Fasy[1], David Millman[1], Samuel Micka[2], Luke Padula[3], Maksym Makarchuk[1]

[1]Montana State University, [2]Western Colorado State University, [3]Colorado State University

Abstract

Topological descriptors such as the Euler characteristic curve and persistence diagrams are increasingly used to represent complex data. Recent research indicates that a carefully chosen set of these descriptors can capture both geometric and topological information about shapes in d-dimensional space. In practical applications, epsilon-nets are employed for data sampling, presenting two extremes: oversampling, where a small epsilon ensures a comprehensive representation but may lead to computational inefficiencies, and undersampling, where epsilon lacks a grounded rationale, offering faster computations but risking an incomplete shape description without theoretical guarantees. This study investigates the phenomena of oversampling and undersampling across synthetic and real-world datasets.

Background

In this research project, we utilize finite simplicial complexes as our data representation. We assume that the data is in general position, meeting the necessary requirements for our analysis.

Defenition 1: A finite simplicial complex is a finite set of simplices (basic geometric figures such as points, edges, triangles, and n-dimensional simplices)..

Defenition 2: A filtration is a sequence of subcomplexes starting from the empty set to the simplicial complex itself, particularly noting subsets where topological changes take place. An example of a filtration in action is provided in Figure 1.

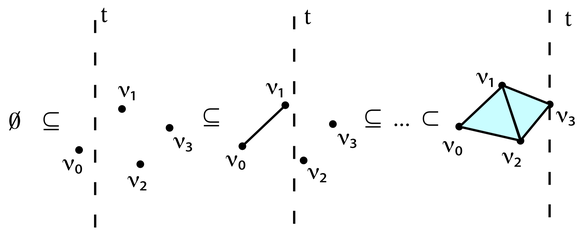

Figure-1
Examples of 0-dimensional, 1-dimensional, 2-dimensional, and 3-dimensional simplices

Defenition 3: A topological descriptor of height filtration is a any invariant of that filtration. [Edelsbrunner and Harer (2010)]. For purpose of this project we will focus on *persistence homology transform* (PHT) and its' finite representation.

Defenition 4: A persistence diagram is a multiset of points in the plane, with each point corresponding to a topological feature (connected component, hole, etc.), such that the coordinates of each point represent the birth and death scales of the corresponding feature, respectively.

Data

In our experiments, we worked with three types of data: random point clouds (RANDPTS), the MPEG7 dataset, and a subset of EMNIST. The RANDPTS dataset consists of varying-sized point clouds generated from a uniform distribution within a $[0, 10]^2$ box. MPEG7 includes 70 classes of binary shape images, while EMNIST contains 62 classes of grayscale handwritten characters. For the MPEG7 and EMNIST datasets, we applied a simple preprocessing pipeline based on contours. First, we convert to a binary image based on a threshold. To compute the global threshold, we use Otsu's (1975) algorithm. Second, for each binary image, we compute one contour curve at two simplification levels as follows. For each binary image, we apply Suzuki and Abe's (1985) algorithm to extract the contour of the image and then apply Douglas–Peucker's (1973) method to simplify the contour.

Experiments

In this study, we are going to implement code for a few experiments describing potential issues of TDA. In Geometric based discretization experiment we'll demonstrate the impact of oversampling directions from a unit sphere on a size of set directions needed to fully represent data. We'll examine how the minimum size stratum for sampling evolves with increasing vertices in the simplicial complex. These findings will be visualized on a log-log scatter plot, and we'll derive the best-fit line to reveal the relationship between the number of vertices in a simplicial complex and the size of the smallest stratum.

In all three experiments, as the number of vertices increases, smaller layers emerge, particularly noticeable for smaller minimal stratum sizes. This observation discourages attempts to discretize the transformation with a minimal observation region, as selecting directions based on data geometry can lead to an excessively large number of detailed directions, posing computational challenges and potentially rendering analysis results meaningless.

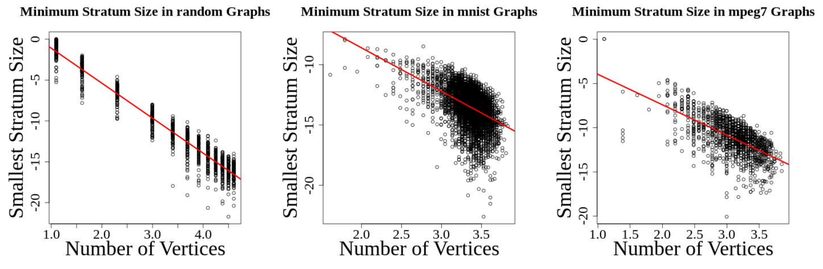

Figure-1
Minimum stratum size graphs for RANDPTS, EMNIST, and MPEG7

In the constant size discretization experiment we are going to fix the size of the set of directions for the filtration function and study the proportion of sampeled subset from unit sphere that hit observeble regions as we vary the number of vertices. In practice, widely common to use a fixed size of the directions set, however this might result in losing important information about the shape. In this experiment, we chose 214 = 16384 directions. We chose such a big number of directions in order to study the best-case scenario.Regrettably, we observe a decrease in the number of hits in the stratification as the number of vertices increases. This outcome was anticipated, given that with each vertex, the geometry of the simplex becomes more intricate. Even when disregarding exceptions, we detect a moderately strong negative correlation.

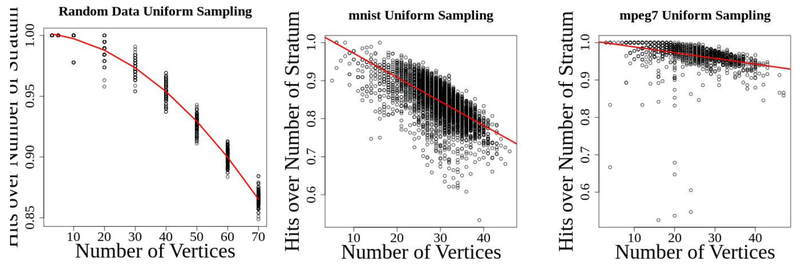

Figure-1
Uniform sampling graphs for RANDPTS, EMNIST, and MPEG7

Two previous experiments have revealed potential issues with conventional methods of sampling directions for TDA analysis. In this experiment, we aim to showcase an alternative approach using real-world data. For the final experiment, we'll employ the Bozeman road network as input in our code. We'll downsample the network into 4, 5, and 6-vertex graphs, identify all possible planar graphs per graph, and group them together. Subsequently, we'll filter the simplicial complexes and determine the number of directions required to identify a graph.

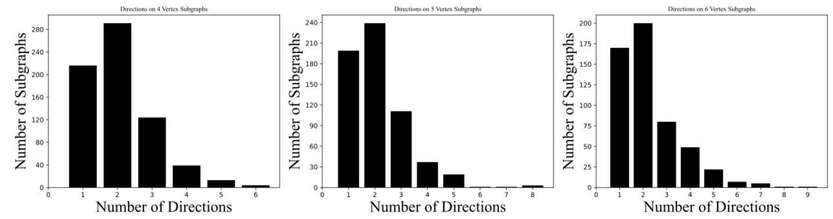

Figure-1
Direction distribution for 4-vertex, 5-vertex, and 6-vertex graphs

As we can see, typically only two directions are needed to identify the original graph. Therefore, using coarse stratification to sample from the unit sphere appears to be an optimal choice. This method simplifies the sampling process while preserving essential graph information, making it both practical and efficient for various applications.

References

Douglas, David H., and Thomas K. Peucker. "Algorithms for the reduction of the number of points required to represent a digitized line or its caricature." Cartographica: the international journal for geographic information and geovisualization 10.2 (1973): 112-122.

Edelsbrunner, Herbert, and John L. Harer. Computational topology: an introduction. American Mathematical Society, 2010.

Otsu, Nobuyuki. "A threshold selection method from gray-level histograms." Automatica 11.285-296 (1975): 23-27.

Suzuki, Satoshi. "Topological structural analysis of digitized binary images by border following." Computer vision, graphics, and image processing 30.1 (1985): 32-46.