INSTABILITY OF THE BETTI SEQUENCE FOR PERSISTENT HOMOLOGY AND A STABILIZED VERSION OF THE BETTI SEQUENCE

MEGAN JOHNSON[1] AND JAE-HUN JUNG[2,†]

[1]DEPARTMENT OF MATHEMATICS, UNIVERSITY AT BUFFALO, THE STATE UNIVERSITY OF NEW YORK, BUFFALO, NY, 14260, USA
Email address: meganjoh@buffalo.edu

[2]DEPARTMENT OF MATHEMATICS & POSTECH MATHEMATICAL INSTITUTE FOR DATA SCIENCE (MINDS), POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY, POHANG 37673, KOREA
Email address: †jung153@postech.ac.kr

ABSTRACT. Topological Data Analysis (TDA), a relatively new field of data analysis, has proved very useful in a variety of applications. The main persistence tool from TDA is persistent homology in which data structure is examined at many scales. Representations of persistent homology include persistence barcodes and persistence diagrams, both of which are not straightforward to reconcile with traditional machine learning algorithms as they are sets of intervals or multisets. The problem of faithfully representing barcodes and persistent diagrams has been pursued along two main avenues: kernel methods and vectorizations. One vectorization is the Betti sequence, or Betti curve, derived from the persistence barcode. While the Betti sequence has been used in classification problems in various applications, to our knowledge, the stability of the sequence has never before been discussed. In this paper we show that the Betti sequence is unstable under the 1-Wasserstein metric with regards to small perturbations in the barcode from which it is calculated. In addition, we propose a novel stabilized version of the Betti sequence based on the Gaussian smoothing seen in the Stable Persistence Bag of Words for persistent homology. We then introduce the normalized cumulative Betti sequence and provide numerical examples that support the main statement of the paper.

1. INTRODUCTION

Topological Data Analysis (TDA) is rising field useful for the analysis of high-dimensional data structure [1]. The popularly used tool in TDA is persistent homology, introduced by Edelsbrunner et al. in 2002 [2], where snapshots of the topological structure of the data set are taken at many different scales and the results are compared from one scale to the next. Singular homology is, in general, hard to compute, and almost impossible to compute for real-world data but when the given topological space is approximated from finitely many points,

simplicial homology, can instead be used as it is easily computable. Persistent homology is obtained by computing simplicial homology at different scales. Previous applications of TDA and persistent homology include 3D shape segmentation [3], astrophysics [4, 5, 6], biology and medicines [7, 8], and neuroscience [9, 10] to name a few.

Popular representations of persistent homology include persistent diagrams and persistent barcodes. Persistent barcodes and persistent diagrams are mathematically equivalent and they demonstrate how the homological structures of the given data change according to scale. Although these are useful for data analysis, they are not necessarily compatible with typical machine learning workflows in their raw forms as they are designed as multisets and collections of intervals, respectively. One way to reconcile these persistent homology representations and machine learning algorithms is to vectorize the persistent diagrams or the persistent barcodes [11]. There are several vectorization methods including the topological vector, persistence vector, and persistence images. These vectorization methods are, in general, more computationally efficient than kernel methods. The construction of the topology vectors and persistence vectors are straightforward and easy to implement. The persistence image is less straightforward (i.e. slower) to compute but delivers better classification results, in general. Another vectorization that we consider in this paper is the Betti sequence [12] which contains the Betti numbers of the homology groups of the simplicial complex built on the data set at all scales of the persistent homology.

In this paper, we prove by example that the Betti sequence is unstable with respect to the 1-Wasserstein distance. In other words, a small change in a persistent diagram leads to a large change in the 1-Wasserstein norm of the Betti Sequence. To our knowledge, the instability of the Betti sequence, although mentioned in [13], has not yet been explicitly shown. In practice, such large change may not be significant if finer filtration intervals are chosen but to remedy the instability in the Betti sequence we propose a new stabilized version and prove its stability. In the stabilization, we adopt a similar Gaussian-smoothing approach as in [14, 15]. In this paper, we show numerical examples that support our statement and show the validation of the proposed stabilization of the Betti sequence.

This paper will be organized as follows: Section 2 will cover the background of persistent homology and its representations and it will cover the definition and instability of the Betti sequence. Section 3 will introduce our proposed stabilization of the Betti sequence and provide a proof of its stability. Section 4 will demonstrate the normalized cumulative Betti sequence on various data sets. Section 5 contains our concluding remarks.

2. Persistent Homology and Betti Sequence

The main persistent homology representations that we consider in this paper are the persistence diagram and the persistence barcode. Both are equivalent and each one can be recovered from the other while each can be used differently in terms of numerical manipulations. Let X be a given topological space. Singular homology describes the homological structure of X with the n-dimensional homology group H_n. The homology group, H_n is defined by $H_n = Ker(\delta_n)/Im(\delta_{n+1})$, which is the quotient group of the kernel, $Ker(\delta_n)$, and image

groups, $Im(\delta_{n+1})$, of the boundary map, $\delta_n\colon C_n(X) \to C_{n-1}(X)$ where $C_n(X)$ is the free abelian group whose basis is the set of singular n-simplices in X. Roughly speaking, the rank of the n-dimensional homology group H_n, called the nth Betti number, indicates how many n-dimensional holes are there in X. This information is useful in understanding the topological structure of X. However, computing H_n is not easy as X is an arbitrary topological space in general. In fact, it is not practical to use for X from real-world applications. Thus, in order to use homological features of X for data analysis, instead of using X directly, we use a point cloud sampled from X.

We use typical building algorithms to obtain the point cloud approximation of X, known as the simplicial complex K. There are various ways of constructing K including the Vietoris-Rips complex, $V(\tau)$ where the non-negative real number τ is known as the filtration parameter. With the given value of τ, $V(\tau)$ is constructed by gluing simplices whose pairwise distance is within τ. However, it is not known which value of τ approximates X best. For this reason, we construct V for various τ, which gives us the notion of *persistence*. The nth homology, H_n corresponding to $V(\tau)$ can be defined similarly as above. Let β_n be the Betti number for H_n. β_0 represents the number of connected components in K and β_n the number of n-dimensional cycles or holes. As we have the natural inclusion of $V(\tau_i) \hookrightarrow V(\tau_j)$, $\tau_i \leq \tau_j$ and a homomorphism $H_n(V(\tau_i)) \to H_n(V(\tau_j))$, we have the relation between β_n versus τ, which generates the graph of the persistent barcodes in the considered n-dimension. On the persistent barcodes, an interval of filtration values corresponding to the same β_n is known as a bar and indicates the n-dimensional hole structure of K. If we call the starting point of each bar the *birth* and the ending point the *death* we can create the persistent diagram multiset by considering all points of the form $(birth, death)$. Vectorizations of persistent diagrams and persistent barcodes and their stability are key aspects that we consider in this paper.

2.1. Definition of the Betti Sequence.

2.1. **Definition of the Betti Sequence.** As explained above, now we consider K instead of X by sampling finitely many distinct points from X. Recall the Betti number, β_n, is the rank of the nth dimensional homology group $H_n(K, R)$. More specifically, β_0 represents the number of connected components in K, β_1 the number of 1-dimensional holes, etc. As we have the natural inclusion of our filtered simplicial complex $\{VR_{\tau_i}\}$, namely $i\colon VR_{\tau_i} \hookrightarrow VR_{\tau_j}$, $\tau_i \leq \tau_j$, we have a homomorphism $H_n(VR_{\tau_i}) \to H_n(VR_{\tau_j})$. This defines a relationship between β_n and τ, which is used to generate the persistence barcode of dimension n.

The Betti sequence, or Betti curve, originally defined in 2017, is the vectorization of Betti numbers obtained in persistent homology [12]. The Betti sequence uses the persistence barcode and a discretization of the filtration interval to define the vectorization. At each value of τ in the discretization, we count the number of generators existing at that filtration and that is our vector entry. We provide the formal definition of the Betti sequence below.

Definition 2.1. Given a persistence barcode of dimension n with finitely many persistence intervals and a maximum filtration $\tau_{max} > 0$, let $\{\tau_i\}_{i=1}^{M}$ be equally spaced points in $[0, \tau_{max}]$. Let $\vec{v}_b = (v_i)_1^M$ be the vector whose entries v_i count the number of persistence intervals in the barcode existing for the filtration value τ_i.

This definition suffers from the following flaw: persistence bars, if they fall entirely in between discretization values, are not counted at all and have no impact on the Betti sequence. Note that if the mesh size is exceedingly small then any bar which falls entirely in between the tau values is likely due to noise and their omission might not be an issue. However, if the mesh size is chosen poorly this could result in a major loss of information.

We propose an alternate definition of the Betti sequence which agrees with the original definition in the limit as the mesh size goes to zero.

Definition 2.2. Given a persistence barcode of dimension n with finitely many bars and a maximum filtration $\tau_{max} > 0$, divide the interval $[0, \tau_{max}]$ into M equal subintervals of length $\Delta\tau = \frac{\tau_{max}}{M}$. Let $\vec{v}_b = (v_i)_1^M$ be the vector whose entries v_i count the number of bars in the barcode that exist for at least one filtration value τ in the ith subinterval (τ_{i-1}, τ_i) of the filtration interval $[0, \tau_{max}]$.

This alternate definition has the added benefit of being easily translatable to the language of persistence diagrams making the study of the stability of the Betti sequence with respect to the p-Wasserstein metric possible.

2.2. **Redefining the Betti Sequence via the Persistence Diagram.** In order to discuss stability with respect to the Wasserstein metric, we need to redefine the Betti sequence in terms of the persistent diagram. Consider a persistent diagram B with finitely many off-diagonal points. Then a bar on the corresponding persistent barcodes exists at some filtration value τ in the subinterval (τ_{i-1}, τ_i) if and only if its birth-death point on the persistent diagram falls in the shaded region, call it D_i, illustrated in Figures 1 and 2
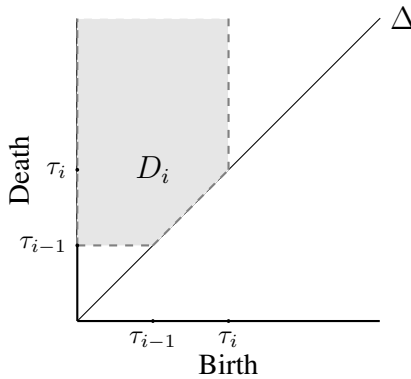

FIGURE 1. The shaded region, D_i, of the persistence diagram corresponding to the filtration interval $[\tau_{i-1}, \tau_i]$ on a barcode. Here Δ is the diagonal.

More precisely, a bar on a barcode exists at some filtration value τ in the subinterval (τ_{i-1}, τ_i) if and only if its birth-death point on the persistence diagram falls in one of the following four subregions of D_i:
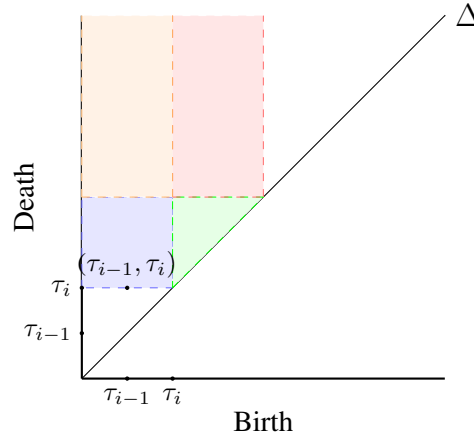
FIGURE 2. The shaded region, D_i, of the persistence diagram corresponding to the filtration interval $[\tau_{i+1}, \tau_i]$ on a barcode for four different cases. Here Δ is the diagonal.

- If the birth-death point lies in the red region, then the bar begins after τ_{i-1} and ends after τ_i.
- If the birth-death point lies in the orange region, then the bar begins before τ_{i-1} and ends after τ_i.
- If the birth-death point lies in the green region, then the bar begins after τ_{i-1} and ends before τ_i.
- If the birth-death point lies in the blue region, then the bar begins before τ_{i-1} and ends before τ_i.

Let us now make a precise description of the D_i as a subset of $\mathbb{R}^2$ (with multiplicity). Let $D_i = K_i \setminus C_i$ where

$$K_i = \left\{ (b,d) \in \mathbb{R}^2 \mid \quad 0 < b < \tau_i, d > \tau_{i-1} \right\}$$

and

$$C_i = \left\{ (b,d) \in \mathbb{R}^2 \mid \quad \tau_{i-1} \le b \le \tau_i, \tau_{i-1} < d < b \right\}$$

Definition 2.3. Let B be a persistence diagram with finitely many off-diagonal points. The Betti sequence of B is defined as

$$\vec{v}(B) = (v_i)_1^N \in \mathbb{R}^N$$

where $v_i = |D_i \bigcap B|$ is the cardinality of the intersection of D_i, described above, and the persistence diagram B.

2.3. **Instability of the Betti Sequence.** Recall the definition of the p-Wasserstein distance between persistence diagrams [16].

Definition 2.4. The p-Wasserstein distance between persistence diagrams X and Y is

$$W_p(X, Y) = \left[\inf_{\eta : X \to Y} \sum_{x \in X} ||x - \eta(x)||_\infty^p \right]^{1/p}$$

where $\eta : X \to Y$ is a partial matching of X and Y. Note as $p \to \infty$ this distance becomes the *bottleneck distance*.

Theorem 2.5. *Let B and B' be persistence diagrams containing only finitely many off-diagonal points. The Betti sequence is **not stable** with respect to the 1-Wasserstein distance. That is, there exists persistence diagrams B and B' such that there does not exist a non-negative constant C such that*

$$||\vec{v}(B) - \vec{v}(B')||_\infty \leq C \cdot W_1(B, B')$$

Proof. We will prove by example. Let B be a persistence diagram with finitely many off-diagonal points and with maximum filtration τ_{max}. Suppose further, for diagram B, that the number of subintervals, N, of $[0, \tau_{max}]$ is fixed so that there exists exactly one birth-death point $x_i = (b_i, d_i)$ with $\tau_{i-1} < b_i, d_i < \tau_i$ in each of the non-overlapping parts of the regions D_i, described above, for $1 \leq i \leq N$ as seen in Fig. 3. Fix an index j and for any $\epsilon > 0$ let $d_j = \tau_j - \frac{\epsilon}{2}$. The Betti sequence vector, $\vec{v}(B)$, of this persistence diagram is then, by definition, given by the following

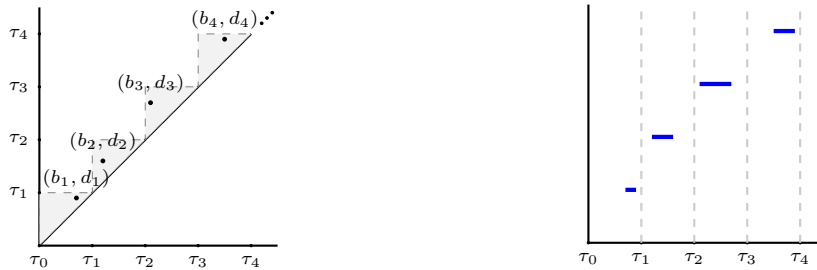$$\vec{v}(B) = \langle 1, ..., 1, 1, 1, ..., 1 \rangle^T$$

FIGURE 3. Left: The persistence diagram where all persistence points exist in the non-overlapping regions of the D_i (the shaded triangles). Right: The corresponding persistence barcode.

Now consider another persistence diagram B' with finitely many off-diagonal points with the same maximum filtration, τ_{max} and the same number of intervals, N. Suppose B' is almost an exact copy of B except that d_j has been shifted by ϵ to become $d_j' = d_j + \epsilon = \tau_j + \frac{\epsilon}{2}$. Then the persistence point $x_j = (b_j, d_j')$ is in D_{j+1} and the Betti sequence for B' is

$$\vec{v}(B') = \langle 1, ..., 1, 2, 1, ..., 1 \rangle^T$$

where $v_j = 1, v_{j+1} = 2$, and $v_{j+2} = 1$. For the Betti sequence vectorization to be stable with respect to the 1-Wasserstein distance under the small perturbation of ϵ we need a non-negative constant C such that

$$||\vec{v}(B) - \vec{v}(B')||_\infty \leq C \cdot W_1(B, B')$$

Clearly, $||\vec{v}(B) - \vec{v}(B')||_\infty = 1$ and if we recall the definition of the 1-Wasserstein distance

$$W_1(B, B') = \inf_{\eta:B \to B'} \sum_{x \in B} ||x - \eta(x)||_\infty$$

where η is a partial matching of B and B', we know that $W_1(B, B') = ||(b_j, d_j) - (b_j, d'_j)||_\infty = \epsilon$. Thus for stability we need a non-negative constant C such that $1 \leq C \cdot \epsilon \iff \frac{1}{\epsilon} \leq C$. However, as $\epsilon > 0$ can be made arbitrarily small, there does not exist such a constant C. Therefore the Betti sequence is unstable with respect to the 1-Wasserstein distance. $\qquad \square$

Remark 2.6. *Similarly, it can be shown that the Betti sequence is unstable with respect to the Wasserstein distance with $p \to \infty$, i.e. with respect to the bottleneck distance.*

3. STABILIZED BETTI SEQUENCE

We now propose a stabilized version of the Betti sequence inspired by the Gaussian smoothing techniques seen in [14, 15] and prove its stability with respect to the 1-Wasserstein distance.

Definition 3.1. Suppose B is a persistence diagram with M off-diagonal points $x_j = (b_j, d_j)$ and maximum filtration τ_{max}. Divide the interval $[0, \tau_{max}]$ into N equal subintervals of the form $[\tau_{i-1}, \tau_i]$ each of length $\Delta\tau$ as above. Let $\{G_i; \mu_i, \Sigma_i\}_{i=1}^N$ be a collection of Gaussian distributions where μ_i, the mean of G_i, is chosen to be $\langle \tau_{i-1}, \tau_i \rangle^T$ and where Σ_i is the covariance matrix for G_i. Define w_i to be $w_i = \frac{1}{N}$ and note that each $w_i > 0$ and $\sum_{i=1}^N w_i = 1$. Then the stable Betti Sequence vector is defined by

$$\vec{v}^{\,s} = \left(v_i^s = w_i \sum_{j=1}^M p_i(x_j \,|\, \mu_i, \Sigma_i) \right)_{i=1}^N$$

where

$$
\begin{aligned}
p_i(x_j \,|\, \mu_i, \Sigma_i) &= \frac{\exp\left(-\frac{1}{2}(x_j - \mu_i)^T \Sigma_i^{-1}(x_j - \mu_i)\right)}{2\pi |\Sigma_i|^{1/2}} \\
&= \frac{\exp\left(-\frac{1}{2\Delta\tau}||x_j - \mu_i||_2\right)}{2\pi \Delta\tau}
\end{aligned}
$$

is the probability density function of Gaussian G_i at x_j. Note that $|\Sigma_i|$ is the determinant of the covariance matrix, Σ_i.

Remark 3.2. *The choice of Σ_i is still an open question. Ideally, we would want to use a sharp Gaussian and so Σ_i should be defined so that $p_i(x_j \,|\, \mu_i, \Sigma_i)$ is essentially the same for every persistence point x_j in D_i not "near" the boundary of D_i. The choice of Σ_i should be further studied in future work.*

Theorem 3.3. *Let B be a persistence diagram of finite size and B' be the persistence diagram obtained by perturbing B by an arbitrary $\epsilon > 0$ such that $W_1(B, B') \leq \epsilon$. Then there exists a non-negative constant $C < \infty$ for any ϵ such that*

$$||\vec{v}^{\,s}(B) - \vec{v}^{\,s}(B')||_\infty \leq C \cdot W_1(B, B')$$

Proof. Let M be the number of off-diagonal points in B. Let $\eta : B \to B'$ be the partial matching that realizes the 1-Wasserstein distance between B and B'. For a fixed $i \in \{1, ..., N\}$ we have

$$\left| v_i^{s}(B) - v_i^{s}(B') \right| = \left| w_i \sum_{j=1}^{M} \left(p_i(x_j \,|\, \mu_i, \Sigma_i) - p_i(\eta(x_j) \,|\, \mu_i, \Sigma_i) \right) \right|$$

$$\leq w_i \sum_{j=1}^{M} \left| p_i(x_j \,|\, \mu_i, \Sigma_i) - p_i(\eta(x_j) \,|\, \mu_i, \Sigma_i) \right|$$

As $p_i : \mathbb{R}^2 \to \mathbb{R}$ is continuously differentiable it is also Lipschitz continuous with Lipschitz constant L_i. We get

$$w_i \sum_{j=1}^{M} \left| p_i(x_j \,|\, \mu_i, \Sigma_i) - p_i(\eta(x_j) \,|\, \mu_i, \Sigma_i) \right| \leq w_i \sum_{j=1}^{M} \left| L_i(x_j - \eta(x_j)) \right|$$

$$= w_i L_i \sum_{j=1}^{M} \left| x_j - \eta(x_j) \right|$$

$$= w_i L_i \cdot W_1(B, B')$$

If we let $C = \max_{i \in [1, ..., N]} w_i L_i$ we have the desired result. $\square$

Example 3.4. Returning to a simplified version of the example used to show that the Betti sequence was unstable, we will now show that Theorem 3.3 is satisfied for this example.

Let B be the persistence diagram that contains two off-diagonal points: $x_1 = \left(0.25, 0.5 - \frac{\epsilon}{2}\right)$ and $x_2 = (0.75, 0.8)$ (see Fig. 4) for any $0 < \epsilon$.

The stable Betti sequence for the persistence diagram B, using maximum filtration 1 and two subintervals, is easily computed as the vector $\vec{v}^{S}(B) = \langle v_1^S, v_2^S \rangle^T$ where

$$v_1^S = \frac{1}{2\pi} \left[\exp\left(-\frac{1}{16} - \frac{\epsilon^2}{4} \right) + \exp\left(-\frac{549}{400} \right) \right]$$

$$v_2^S = \frac{1}{2\pi} \left[\exp\left(-\frac{1}{16} - \frac{\epsilon^2}{4} \right) + \exp\left(-\frac{41}{400} \right) \right]$$

Now let B' be the persistence diagram, pictured in Fig. 5, that contains two off-diagonal points: $y_1 = \left(0.25, 0.5 + \frac{\epsilon}{2}\right)$ and $y_2 = (0.75, 0.8)$, for any $0 < \epsilon$.
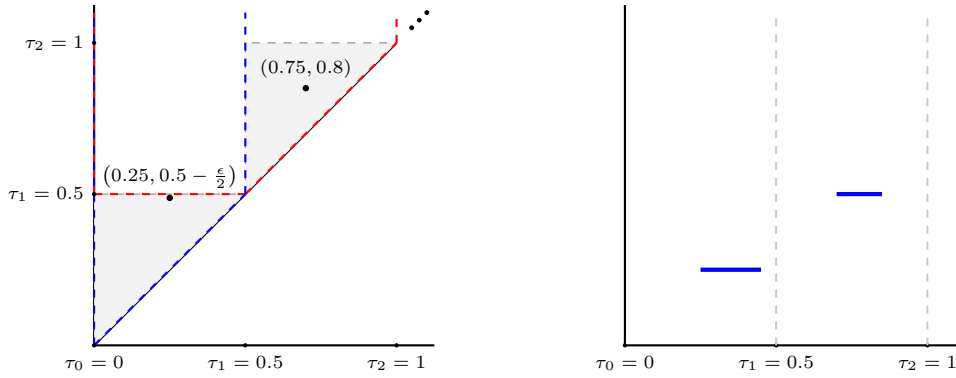
FIGURE 4. Left, the persistence diagram B with two persistence points $\left(0.25, 0.5 - \frac{\epsilon}{2}\right)$ and $(0.75, 0.8)$. The dashed blue lines outline the region D_1 used in the definition of the Betti sequence and the red dashed lines outline the region D_2. Right, its corresponding persistence barcode.
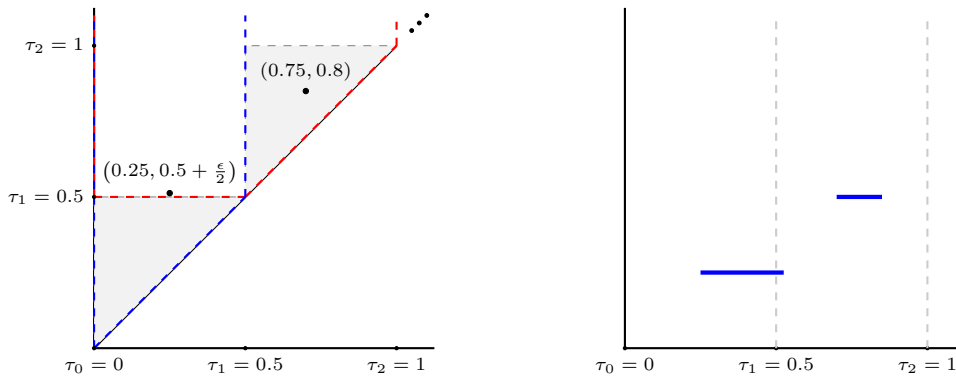

FIGURE 5. Left, the persistence diagram B' with two persistence points $\left(0.25, 0.5 + \frac{\epsilon}{2}\right)$ and $(0.75, 0.8)$. The dashed blue lines outline the region D_1 used in the definition of the Betti sequence and the red dashed lines outline the region D_2. Right, its corresponding persistence barcode.

The stable Betti sequence for the persistence diagram B', using maximum filtration 1 and two subintervals, is also easily computed as the vector $\vec{v}^S(B') = \langle v_1^S, v_2^S \rangle^T$ where

$$v_1^S = \frac{1}{2\pi} \left[\exp\left(-\frac{1}{16} - \frac{\epsilon^2}{4}\right) + \exp\left(-\frac{549}{400}\right) \right]$$

$$v_2^S = \frac{1}{2\pi} \left[\exp\left(-\frac{5}{16} + \frac{\epsilon}{2} - \frac{\epsilon^2}{4}\right) + \exp\left(-\frac{41}{400}\right) \right]$$

Thus, we obtain

$$\vec{v}^S(B) - \vec{v}^S(B') = \begin{bmatrix} 0 \\ \frac{1}{2\pi}\left[\exp\left(-\frac{1}{16} - \frac{\epsilon^2}{4}\right) - \exp\left(-\frac{5}{16} + \frac{\epsilon}{2} - \frac{\epsilon^2}{4}\right)\right] \end{bmatrix}$$
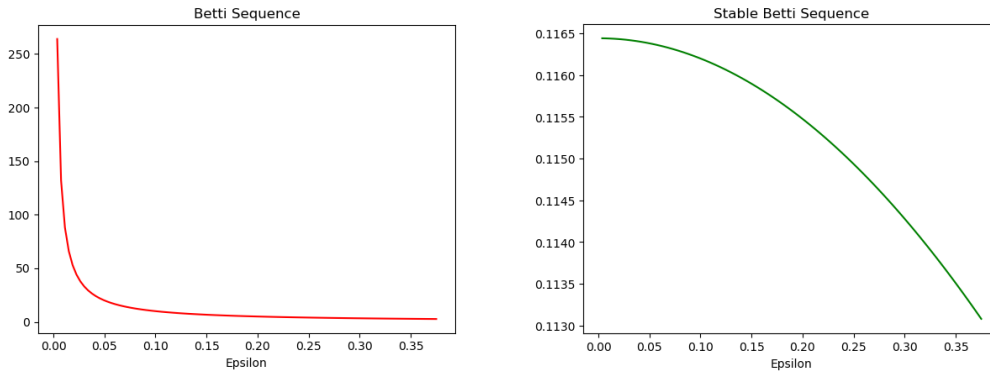
Note that for all $0 < \epsilon$, the absolute value of the first entry in the vector above is less than or equal to that of the second entry in the vector. Thus we obtain

$$||\vec{v} - \vec{v}'||_\infty = \frac{1}{2\pi}\left[\exp\left(-\frac{1}{16} - \frac{\epsilon^2}{4}\right) - \exp\left(-\frac{5}{16} + \frac{\epsilon}{2} - \frac{\epsilon^2}{4}\right)\right]$$

FIGURE 6. The plot of the ratio of the L_∞-norm of the change in the original Betti sequence (left) and the stable Betti sequence (right) to the 1-Wasserstein distance between persistence diagrams B and B'.

Recall, in order to satisfy Theorem 3.3 we need

$$||\vec{v}^S(B) - \vec{v}^S(B')||_\infty \leq C \cdot W_1(B, B')$$
$$= C||\left(0.25, 0.5 - \frac{\epsilon}{2}\right) - \left(0.25, 0.5 + \frac{\epsilon}{2}\right)||_\infty$$
$$= C\epsilon$$

This means that we need to find a constant C such that

$$\frac{1}{\epsilon} \cdot ||\vec{v}^S(B) - \vec{v}^S(B')||_\infty \leq C$$

The right hand plot in Fig. 6 contains the graph of the left hand side of the inequality above versus ϵ. The graph attains a maximum value of approximately 0.116440243790144 as ϵ goes to 0. Thus there exists C, say, $C = 0.2$, so that

$$||\vec{v}^S(B) - \vec{v}^S(B')||_\infty \leq C\epsilon$$

and Theorem 3.3 is satisfied. Note that for this example, the L_∞ change in the original Betti sequence is 1 and the 1-Wasserstein distance between the persistence diagrams is ϵ. Therefore, as evidenced in Fig. 6, there is no maximum value of the ratio of the L_∞-norm to the 1-Wasserstein distance for the original Betti sequence.

4. NUMERICAL EXAMPLE

For the numerical experiments, we first consider four point clouds: a uniform lattice of points with a small uniform random coordinate perturbation of magnitude at most $1/N$ on $[0,1]^2$, a uniformly random distribution on $[0,1]^2$, points drawn from the Sierpinski triangle created by the *chaos game* [17] and a uniformly random distribution with a square hole of $[0.15, 0.85]^2$ on $[0,1]^2$. A sample point cloud for each type is shown in Fig. 7.
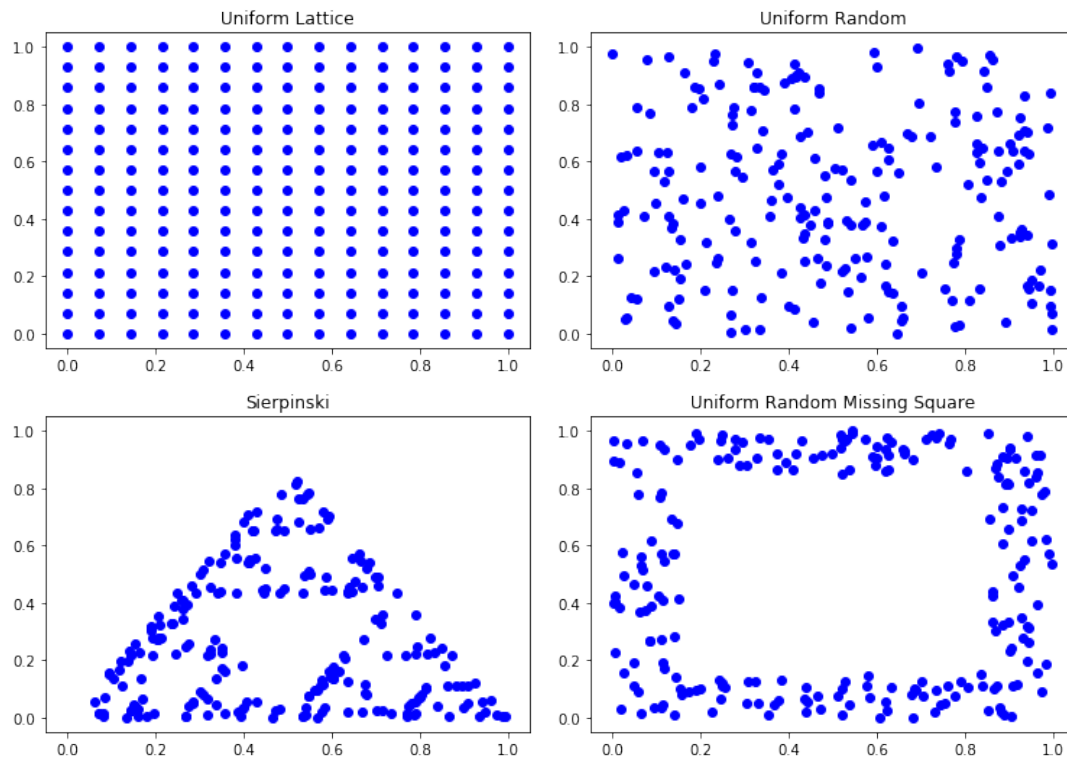

FIGURE 7. Top: a uniform lattice of points with a small uniform random coordinate perturbation of magnitude at most $1/N$ and a uniform random distribution on $[0,1]^2$. Bottom: points drawn from the Sierpinski triangle created by the chaos game and a uniformly random distribution with a square hole of $[0.15, 0.85]^2$ removed on $[0,1]^2$.

Figure 8 shows a sample persistence barcode and diagram for the uniform random data (left) and Sierpinski data (right) for both zero-dimensional (red) and one-dimensional homology (blue). We observe the difference in each barcode and diagram between the uniform and Sierpinski data. Here we note that the Sierpinski data is generated with the uniform random sampling within the chaos game.
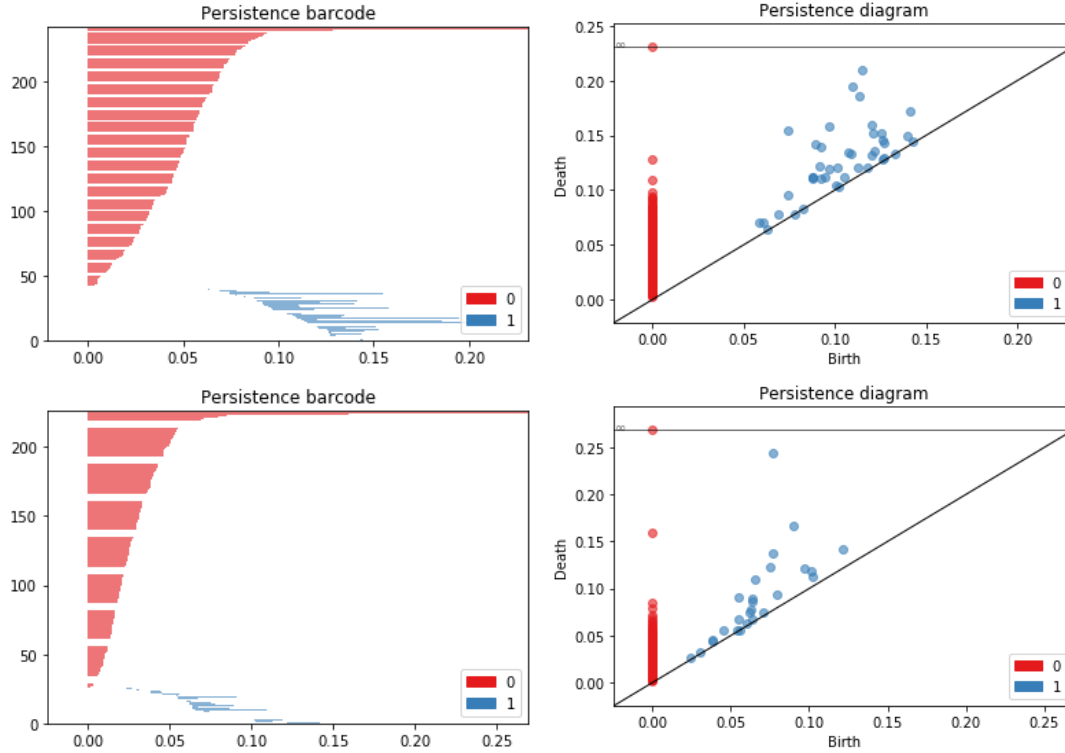
FIGURE 8. Persistence barcode and diagram. Top: Uniform data. Bottom: Sierpinski data.

We will show the instability of the original Betti sequence and compare the results with the stable Betti sequence. For the stable Betti sequence, in the absence of an ideal covariance matrix Σ_i for each Gaussian G_i, we use the adapted Gaussian-smoothing approach as defined below: Consider the following set, X_i

$$X_i = \{x_j | \tau_i < b_j < \tau_i + \gamma \Delta \tau \} \cup \{x_j | \tau_{i-1} - \gamma \Delta \tau < d_j < \tau_{i-1}\}$$

where $\gamma > 0$ is a positive constant. Since the instability we described above is induced near the domain boundary, we consider points outside D_i but near the boundary of D_i. For the numerical example, we consider a sharp Gaussian such that all the points in $D_i \cup X_i$ participate in the Betti sequence. Since the instability is more sensitive to the lower indices of the Betti sequence, we choose the free parameter γ as below (notice that if $\gamma = 0$, it reduces to the original Betti sequence)

$$\gamma = (N - i + 1)/10, \quad i = 1, \cdots, N$$

Let $\tilde{D}_i$ be $X_i \cup D_i$ and $|\tilde{D}_i|$ be the cardinality of $\tilde{D}_i$. Then define a new stable Betti sequence $v^{new}(B)$ as

$$\vec{v}^{new}(B) = (|\tilde{D}_1|, |\tilde{D}_2|, \cdots, |\tilde{D}_N|)^T$$

$\vec{v}^{new}(B)$ is the original Betti sequence with each domain extended by $\gamma \Delta \tau$, which can also be viewed as the Betti sequence Gaussian-smoothed with a sharp truncation near the domain boundary. Further, we define the cumulative Betti sequence, $\vec{v}^{cum}(B)$ of $\vec{v}(B)$ recursively as

$$\vec{v}_1^{cum}(B) = \vec{v}_1(B)$$
$$\vec{v}_i^{cum}(B) = \vec{v}_i(B) + \vec{v}_{i-1}^{cum}(B), \quad i \geq 2$$

Then the normalized cumulative vector $\hat{v}^{cum}(B)$ is defined as

$$\hat{v}^{cum}(B) = \frac{\vec{v}^{cum}(B)}{||\vec{v}^{cum}(B)||_\infty}$$

We first show the instability of the Betti sequence. Consider the interval, $(x, y) \in \Omega = [0, L_x] \times [0, L_y] := [0, 1 + \epsilon] \times [0, 1 + \epsilon]$, where $\epsilon \ll 1$. We consider a uniform lattice of points with the total number of bins for the Betti sequence, $N = 20$, and the total number of data points, 225. We rescale the lattice by mapping Ω to $\Omega \times 14/N$ while keeping the filtration interval $\Delta \tau = 1/N$. For this case, the shortest lattice interval becomes the same as the filtration interval $\Delta \tau = 1/N = 1/20$ when $\epsilon = 0$ and the corresponding normalized cumulative Betti sequence becomes

$$\hat{v}^{cum}(B) = < 0, 1, 1, \cdots, 1 >$$

because the shortest lattice interval coincides with the domain interval, $\Delta \tau$, in the persistent diagram. However, if $\epsilon < 0$, the shortest interval becomes less than $\Delta \tau$ and so the corresponding Betti sequence is

$$\hat{v}^{cum}(B) = < 0.5, 1, 1, \cdots, 1 >$$

Here note that if $\epsilon > 0$ the birth of all the bars in the barcode remain in the second domain and the death of all the bars still remain in the second domain because the perturbation ϵ is chosen small enough. Thus if the perturbation ϵ is small enough, the first element of the Betti sequence has the value of 0 or 0.5 while the same element in the stable Betti sequence remains almost same under the small perturbation ϵ. The left figure in Fig. 9 shows the first element of the Betti sequence versus ϵ, blue for the Betti sequence and red for the stable Betti sequence with the perturbation $\epsilon \in (-10^{-8}, 10^8) - 100$ values of ϵ were chosen uniformly. The right figure shows the same plot for the uniform random data whose domain is also $[0, 1 + \epsilon] \times [0, 1 + \epsilon]$. As shown in the figure, the original Betti sequence fluctuated between 0 and 0.5 as expected in the left figure and it also fluctuates more than the stable Betti sequence in the right figure. Now in Fig. 10 we show both original Betti sequence (blue) and stable Betti sequence (red) for the four cases shown in Fig. 7 with the fixed domain size $[0, 1] \times [0, 1]$ and $N = 20$.

We use 100 samples for each case and plot all the Betti sequences. The figure shows that the stable Betti sequences yield more sharp patterns while maintaining similar vector structure overall. In addition, in each data set the stable Betti sequences are more homogeneous than the Betti sequences which is promising with regards to possible machine learning applications.
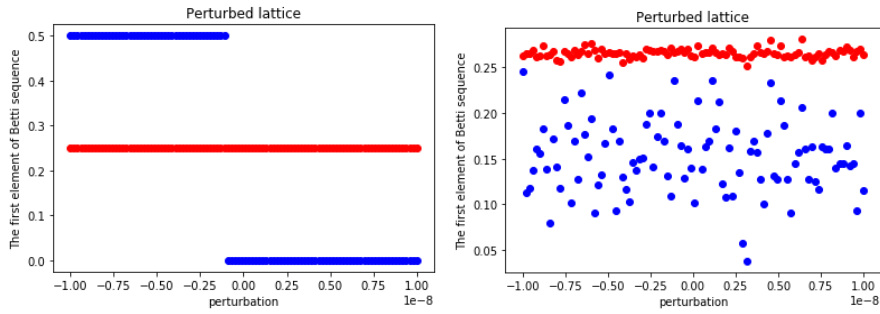
FIGURE 9. v_1, the first element in the sequence, versus ϵ. Blue: Betti sequence. Red: Stable Betti sequence. Left: Perturbed lattice with ϵ. Right: Uniform random data.
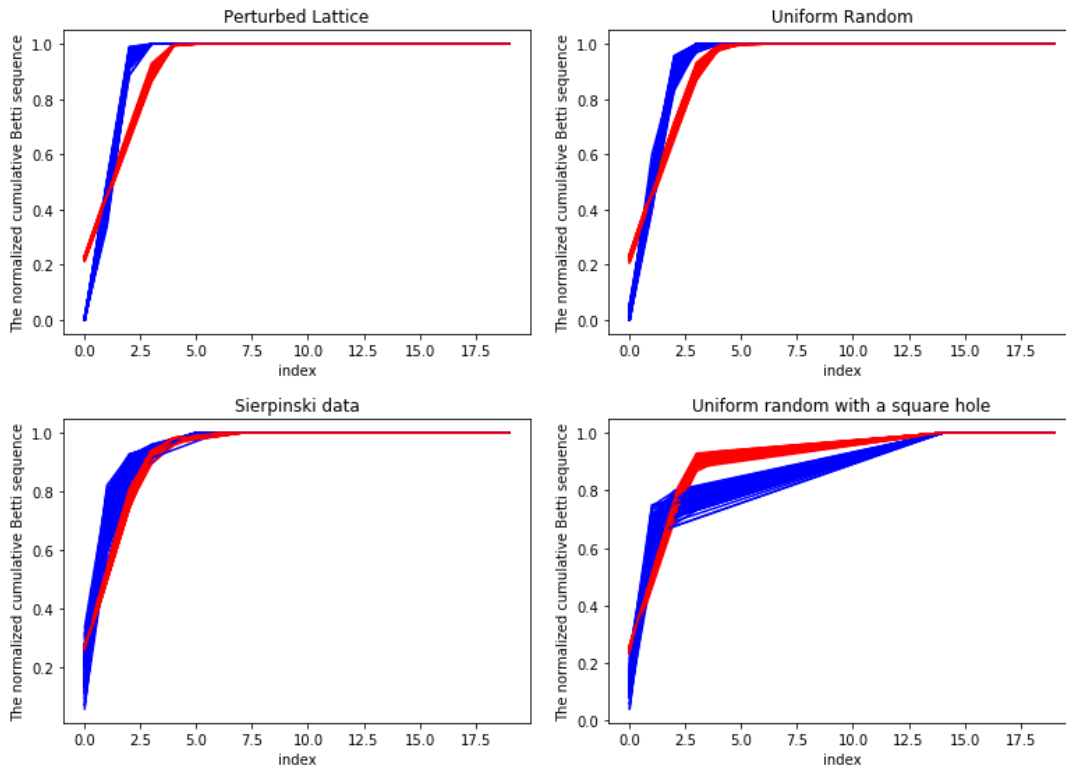

FIGURE 10. The Betti sequence (blue) and stable Betti sequence (red) with fixed domain size, $[0,1] \times [0,1]$. Top: a lattice with a small perturbation and a uniform random data. Bottom: a Sierpinski data and a uniform random with a square hole. Each point sequence was computed with $N = 20$ and there were 225 points in each data set.

5. Concluding Remarks

Topological data analysis and its main tool, persistent homology, has recently gained attention in the scientific community and has proven to be highly useful in various applications. Recently, sizable research has been conducted to combine topological data analysis and machine learning. However, the representations of persistent homology, the persistence diagram and barcode, in their raw forms are not suitable to incorporate into a machine learning workflow and proper feature maps are necessary including vectorization methods. In this paper, we considered the Betti sequence, as a vectorization and showed by example its instability with respect to the 1-Wasserstein distance. In addition, we proposed a stable Betti sequence and proved its stability. With numerical examples, we devised a cumulative stable Betti sequence and showed that the stable Betti sequence was able to achieve a faithful representation of the Betti sequence in that it performs better with the smaller number of the filtration intervals at distinguishing data sets. Our future research will incorporate the proposed stable Betti sequence into machine learning algorithms to study its effectiveness in various applications.

Acknowledgments

References

[1] Gunnar Carlsson. Topology and data. *Bulletin of The American Mathematical Society*, 46:255–308, 04 2009.

[2] Edelsbrunner, Letscher, and Zomorodian. Topological persistence and simplification. *Discrete & Computational Geometry*, 28(4):511–533, Nov 2002.

[3] Mathieu Carrière, Steve Y. Oudot, and Maks Ovsjanikov. Stable topological signatures for points on 3d shapes. *Computer Graphics Forum*, 34(5):1–12, 2015.

[4] Alex Cole and Gary Shiu. Persistent homology and non-gaussianity. *arXiv preprint arXiv:1712.08159*, 2017.

[5] Sven Heydenreich, Benjamin Brück, and Joachim Harnois-Déraps. Persistent homology in cosmic shear: constraining parameters with topological data analysis. *arXiv preprint arXiv:2007.13724*, 2020.

[6] X. Xu, J. Cisewski-Kehe, S.B. Green, and D. Nagai. Finding cosmic voids and filament loops using topological data analysis. *Astronomy and Computing*, 27:34 – 52, 2019.

[7] Melissa R. McGuirl, Alexandria Volkening, and Björn Sandstede. Topological data analysis of zebrafish patterns. *Proceedings of the National Academy of Sciences*, 117(10):5113–5124, 2020.

[8] John Nicponski and Jae-Hun Jung. Topological data analysis of vascular disease: I a theoretical framework. *Frontiers in Applied Mathematics and Statistics*, 6:34, 2020.

[9] Paul Bendich, J. S. Marron, Ezra Miller, Alex Pieloch, and Sean Skwerer. Persistent homology analysis of brain artery trees. *Ann. Appl. Stat.*, 10(1):198–218, 03 2016.

[10] Ann E. Sizemore, Jennifer E. Phillips-Cremins, Robert Ghrist, and Danielle S. Bassett. The importance of the whole: Topological data analysis for the network neuroscientist. *Network Neuroscience*, 3(3):656–673, 2019.

[11] Henry Adams, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier. Persistence images: A stable vector representation of persistent homology. *Journal of Machine Learning Research*, 18(8):1–35, 2017.

[12] Yuhei Umeda. Time series classification via topological data analysis. *Information and Media Technologies*, 12:228–239, 2017.

[13] Yu-Min Chung and Austin Lawson. Persistence curves: A canonical framework for summarizing persistence diagrams. 2020.

[14] Megan Johnson and Jae-Hun Jung. The interconnectivity vector: A finite-dimensional representation of persistent homology. *arXiv preprint arXiv:2011.11579*, 2020.

[15] Bartosz Zieliński, Michał Lipiński, Mateusz Juda, Matthias Zeppelzauer, and Paweł Dłotko. Persistence codebooks for topological data analysis. *Artificial Intelligence Review*, Sep 2020.

[16] L. N. Vaserstein. Markov processes over denumerable products of spaces, describing large systems of automata. *Problems Inform. Transmission*, 5(3):47–52, 1969.

[17] Michael Barnsley. *Fractals everywhere*. Academic Press Inc., Boston, MA, 1988.