Means and medians of sets of persistence diagrams

Katharine Turner

May 11, 2014

Abstract

The persistence diagram is the fundamental object in topological data analysis. It inherits the stochastic variability of the data we use as input. As such we need to understand how to perform statistics on the space of persistence diagrams. This paper looks at the space of persistence diagrams under a variety of different metrics which are analogous to $L^p$ metrics on the space of functions. Using these metrics we can form different cost functions defining different central tendencies and their corresponding measures of variability. This gives us the natural definitions of both the mean and median of a finite number of persistence diagrams. We give a characterization of the mean and the median of an odd number of persistence diagrams. Although we have examples of the mean not being unique nor continuous we prove that generically the mean of sets of persistence diagrams with finitely many off diagonal points is unique. In comparison the sets of persistence diagrams with finitely many off diagonal points which do not have a unique median is of positive measure.

1 Introduction

Topological data analysis is first and foremost meant to be the about the analysis of data. This means that the input is stochastically generated and that we should consider the observed outputs - the persistence diagrams that we produce - as stochastically generated. This introduces the need to understand statistics on the space of persistence diagrams.

Although persistence diagrams have been mainly used as a heuristic there has been a recent surge of interest in analyzing them statistically. There has been great progress in terms of understanding the mean of distributions of persistence diagrams. [9] showed that the space of persistence diagrams is a Polish space and hence it is possible to define (Fréchet) means and showed that such means always exist. These means were then characterized in [12]. Unfortunately, as will be shown in section 4, the mean is neither continuous nor unique. A possible solution is argued in [10] where a probabilistic alternative definition is offered which combines the definition of the mean with the concept of a shaking hand equilibrium in game theory. This probabilistic mean is continuous and unique.

Statistical approaches to topological data analysis are also considered in [5] which explores the distributions of diagrams arising from sampling $S^p$ under various densities. [13] considers a randomization method of null hypothesis testing. Other work relates to the expected bottleneck distances to an observed persistence diagram to the “truth” when considering point clouds approximating a manifold; [6] studies convergence properties and [2] studies confidence
intervals. Alternative approaches to performing statistics is to project the persistence diagrams onto anything space which is easier to analysis in. Examples of this approach include [4] and [3].

Suppose we are given persistence diagrams $X_1, X_2, \ldots, X_N$ from many realizations of point cloud data obtained from the same geometric object. Instead of analyzing any single diagram $X_i$ we may wish to summarize them. We can try to find some central tendencies such as the mean or the median and measure how variable the different persistence diagrams are for the different realizations of point clouds.

Central tendencies are solutions that optimize particular cost functions. These cost functions are based on the $p$–Wasserstein metrics for various $p$. For $p = 2$ the cost function $F_2$ is the mean squared error and the corresponding central tendency is the mean. For $p = 1$ the cost function $F_1$ is the mean absolute deviation and the corresponding central tendency is the median. This process is straightforward when the observations are real numbers. The space of persistence diagrams is far more geometrically complicated than the real line. There are a variety of choices of metric in the space of persistence diagrams and we need to be careful about which metric we use in what circumstance.

Let $\mathcal{D}$ denote the space of persistence diagrams. We consider a family of metrics on $\mathcal{D}$, $d_p$ with $p \in [1, \infty]$, which are analogous to both the $p$–Wasserstein metrics on probability spaces and the $L^p$ metrics on the space of functions on a discrete sets. The metric spaces $(\mathcal{D}, d_p)$ are different in character depending on $p$. To further understand how unmanageable these metric spaces are we explore their curvature.

In section 3 we are ready to define the central tendencies of the mean and the median with their corresponding measures of variability. We construct appropriate cost functions using the appropriate metrics on $\mathcal{D}$. These are compared to the corresponding cost functions defining the mean and the median of sets of real numbers. We then define the mean and the median of sets of points in the plane and copies of the diagonal which is similar to the mean and median of points in plane. We then use this characterization of the mean and median of sets of points and copies of the diagonal to characterize the local minimums of the cost functions $F_1$ and $F_2$. These are described in such a way that we can search through them to find the mean and the median.

In the final section we will explore the issues of uniqueness and discontinuity. We give examples of how the mean and the median is neither unique nor continuous. However we can prove that the sets of persistence diagrams with finitely many off diagonal points that does not have a unique mean has measure zero while those that has do not have a unique median has positive measure.

This paper will not be providing all the proofs for the case $p = 2$ (i.e. the mean) and the reader should instead refer to [12] where the details are given in full. The results are quoted for the sake of completeness and to illustrate to parallels to the case $p = 1$ (i.e. the median).

For the sake of clarity we will be restricting ourselves to when the number of diagrams that we are taking the median of is odd. The ideas will still hold in the case where the number of diagrams are even but the results and the proofs are less clean. This is because when we that the median of an even number of real numbers we do not get a single number but instead get an interval between the middle two entries. This will mean that instead of getting a point in the plane we will be getting rectangles of choice in the plane in which the point must lie. We could decide to take the center of the rectangle. The interested reader can redo the proofs for the even case. We have checked that the main ideas still works.
2 The space of persistence diagrams

We first will recall the definition of a persistence diagram and hence also the definition of the space of persistence diagrams. The set up is we are given a filtration \( K = \{ K_r \mid r \in \mathbb{R} \} \) of a countable simplicial complex indexed over the real numbers with \( K_{-\infty} = \emptyset \). We wish to summarize how the topology of the filtration changes over time.

When \( i < j \), the inclusion map \( \iota^{\{i \rightarrow j\}} : K_i \rightarrow K_j \) induces homomorphisms

\[
\iota^{\{i \rightarrow j\}}_k : H_k(K_i) \rightarrow H_k(K_j)
\]

for each dimension \( k \). We say that a homology class \( \alpha \in H_k(K_i) \) is born at time \( i \) (denoted \( b(\alpha) \)) if it is not in the image \( \iota^{\{i' \rightarrow i\}}_k \) for any \( i' < i \). We say that \( \alpha \) dies at time \( j \) (denoted \( d(\alpha) \)) if \( \iota^{\{i \rightarrow j\}}_k(\alpha) = 0 \) but \( \iota^{\{i \rightarrow j'\}}_k(\alpha) \neq 0 \) for \( i < j' < j \). We say that \( \alpha \) is an essential class of \( K \) if it never dies.

We say the homology class \( \alpha \) has persistence \( d(\alpha) - b(\alpha) \).

For each pair \( (i, j) \) with \( i < j \) we can then consider the vector space of \( k \)-dimensional homology classes that are born at time \( i \) and die at time \( j \). Let \( \beta^{(i, j)}_k \) denote the dimension of this space. Similarly let \( \beta^{(i, \infty)}_k \) denote the dimension of the space of essential \( k \)-dimensional homology classes that are born at time \( i \).

Let \( \mathbb{R}^{2+} := \{(i, j) \in \mathbb{R} \times (\mathbb{R} \cup \infty) : i < j \} \) We define the \( k \)-th persistence diagram corresponding to the filtration \( K \) to be the multi-set of points in \( \mathbb{R}^{2+} \) alongside countably infinite copies of the diagonal such that the number of points (counting multiplicity) in \([i, \infty) \times [j, \infty] \) is equal to dimension of the image of \( \iota^{\{i \rightarrow j\}}_k \). That is it is equal to the dimension of the space of \( k \)-dimensional homology classes that are born at or before \( i \) and die at or after \( j \). This is achieved by placing at each \( (i, j) \) a number of points equal to \( \beta^{(i, j)}_k \). The countably many copies of the diagonal play the role of persistent homology classes whose persistence is zero and hence would not otherwise be seen.

An equivalent way to record the persistent homology information is through barcodes [7] where each off diagonal point \((b(\alpha), d(\alpha))\) corresponds to an interval \([b(\alpha), d(\alpha))\). The copies of the diagonal correspond to empty intervals.

We restrict our attention to persistence diagrams such that \( \sum_{\alpha \text{ not essential}} d(\alpha) - b(\alpha) < \infty \). This is automatically true if the persistence diagrams contain finitely many off diagonal points.

2.1 A family of metrics

Let \( D \) denote the space of persistence diagrams. There are many choices of metric on \( D \) just like there are different choices of metric on spaces of functions. We will consider a family of choices which are analogous to \( p \)-Wasserstein distances on the space of measures or \( L^p \) distances on the space of functions on a discrete set.

Let \( X \) and \( Y \) be persistence diagrams. We can consider bijections \( \phi \) between the points and copies of the diagonal in \( X \) and the points and copies of the diagonal in \( Y \). These are the transport plans that we consider. Bijections always exist because there are countably many copies of the diagonal which everything can be paired with.
For \( p \in [1, \infty) \) define

\[
    d_p(X, Y) = \left( \inf_{\phi: X \to Y} \sum_{x \in X} \|x - \phi(x)\|_p^p \right)^{1/p}.
\]

If we take the limit as \( p \to \infty \) we obtain

\[
    d_\infty(X, Y) = \inf_{\phi: X \to Y} \max_{x \in X} \|x - \phi(x)\|_\infty.
\]

We will call a bijection between points optimal for \( d_p \) if it achieves the infimum in the definition of \( d_p \).

There certainly have been choices made here. For example, in theory one could construct a distance of the form \( \inf_{\phi: X \to Y} \left( \sum_{x \in X} \|x - \phi(x)\|_q^q \right)^{1/q} \) with \( p \) and \( q \) different. However, we feel this would not be as clean in theory nor in practice. Notably if \( \phi((a, b)) = (c, d) \) then \( \|(a, b) - \phi((a, b))\|_p^p = |a - c|^p + |b - d|^p \) but if \( q \neq p \) then no such nice depiction exists. Notably we will see later that the mean and the median (corresponding to cases involving \( p = 2 \) and \( p = 1 \)) are relatively easy to compute as they have nice characterizations. If we instead mixed our \( p \) and \( q \) this would be lost. For example if we used \( p = 1 \) and \( q = 2 \) then we would need - at some point - to calculate the geometric median of a set of points in the plane. It has been shown that in general there is no exact algorithm to find the geometric median of a set of \( k \) points in the plane.

The coordinates in the space of persistence diagrams have particular meanings; one is the birth time and one is the death time. They are often infinitesimally independent (even though not globally so). For example, if we have generated our persistence diagram from a point cloud then each persistence class has its birth and death time (infinitesimally) determined by the location of two pairs of points which are often distinct. Whenever these pairs are distinct, moving any of these four points will change either the birth or the death but not both. The distinctness of the treatment of birth and death times as separate qualities may seem more philosophically pleasing to the reader when in the setting of barcodes.

Observe that \((\mathcal{D}, d_p)\) is disconnected for all \( p \) with a connected component for each number of points lying on the line \( \{(i, \infty) : i \in \mathbb{R}\} \). Other observations are that, for the same pair of diagrams, for different values of \( p \) different bijections may be optimal.

Optimal bijections are not necessarily unique. Non-uniqueness can involve only points off the diagonal or may involve the diagonal.

**Proposition 1.** \((\mathcal{D}, d_p)\) is a geodesic space for all \( p \in [1, \infty] \).

**Proof.** Fix a \( p \in [1, \infty] \) and \( X, Y \in \mathcal{D} \) with \( d_p(X, Y) < \infty \). We want to construct a bijection \( \phi \) such that \( d_p(X, Y)^p = \sum_{x \in X} \|x - \phi(x)\|_p^p \).

Let \( \{\phi_i\} \) be a sequence of bijections such that

\[
    \lim_{i \to \infty} \sum_{x \in X} \|x - \phi_i(x)\|_p^p = d_p(X, Y)^p.
\]

Choose some off-diagonal point \( x \in X \). Consider the sequence \( \{\phi_i(x)\} \). There is convergent subsequence \( \{\phi_{i_j}(x)\} \) which converges to either an off-diagonal point or the diagonal. We will
start our construction of the bijection $\phi$ by choosing $\phi(x)$ to be this limit point. This sequence satisfies
\[
\lim_{j \to \infty} \sum_{x \in X} \|x - \phi_{i_j}(x)\|_p^p = d_p(X, Y)^p.
\]
We now replace our original sequence of bijections $\{\phi_i\}$ with the subsequence $\{\phi_{i_j}\}$. In this manner we can determine a choice $\phi(x)$ for each off-diagonal point $x \in X$. Similarly we can determine $\phi^{-1}(y)$ for all off-diagonal points $y \in Y$. Since we are considering subsequences of subsequences of subsequences we have consistency in our choices.

Since there are only countably many points off the diagonal in the diagrams $X$ and $Y$ combined we can find a bijection $\phi : X \to Y$ with $d_p(X, Y)^p = \sum_{x \in X} \|x - \phi(x)\|_p^p$. From this optimal bijection $\phi$ we can construct a geodesic between $X$ and $Y$. Let $X_t$ be the diagram with off diagonal points $\{(1-t)x + t\phi(x) : x \in X\}$. By observation $X_0 = X$, $X_1 = Y$, and $\{X_t\}$ is distance achieving path.

The case for $p = \infty$ is very similar. We instead consider a sequence of bijections $\{\phi_i\}$ such that $\lim_{i \to \infty} \max_{x \in X} \|x - \phi_i(x)\|_\infty = d_\infty(X, Y)$ and proceed in the case $p \in [1, \infty)$ to produce a bijection $\phi$ such that $d_\infty(X, Y) = \max_{x \in X} \|x - \phi(x)\|_\infty$.

Although we know that an optimal bijection exists it is not necessarily unique. Non-uniqueness can occur involving only off-diagonal points or can involve the diagonal.

In Figure 2.1 we see an example of non-uniqueness involving only points away from the diagonal. This example works for every $p \in [1, \infty]$. Because of symmetry, matching the points vertically or horizontally involves the same cost.

We need different examples of non-uniqueness involving the diagonal for different $p$. Suppose we are considering two persistence diagrams $X$ and $Y$ each containing a single off diagonal point $x$ and $y$ respectively. We care about when the optimal transport plan is $\phi(x) = y$ and when the optimal transport plan is $\phi(x) = \Delta$ and $\phi^{-1}(y) = \Delta$. Given the location of $x$ in the plane there are different regions of the plane depending on $p$ such that $\phi(x) = y$ is the optimal transport plan. These are illustrated in Figure 2 and Figure 3.

Given two diagrams, $X$ and $Y$ each with only finitely many off diagonal points we can find an optimal bijection for $d_p$ using Munkres assignment algorithm (also known as the Hungarian
Figure 2: Case of $p = 1$. Given diagram (red), there region which distinguishes whether it costs less to pair both points to the diagonal (purple) than pairing them to each other (blue). It costs the same on the boundary (green).

Figure 3: Case of $p = 2$. Given diagram (red), there is a parabola which bounds the region which distinguishes whether it costs less to pair both points to the diagonal than pairing them to each other.
algorithm). Munkres algorithm finds the least cost assignment of tasks to people given that there are the same number of tasks as people and each person must be assigned exactly one task. The input is the cost for each person to do each of the tasks. Suppose \( X \) has \( n \) off-diagonal points, labelled \( x_1, x_2, \ldots, x_n \), and \( Y \) has \( m \) off-diagonal points, labelled \( y_1, y_2, \ldots, y_m \). Let \( x_{n+1}, x_{n+2}, \ldots, x_{n+m} \) and \( y_{m+1}, y_{m+2}, \ldots, y_{n+m} \) be copies of the diagonal. We can think of the points and copies of the diagonal in \( X \) as the people and the points and copies of the diagonal in \( Y \) as tasks. The cost of \( x \in X \) doing task \( y \in Y \) is \( \|x - y\|^p \). We construct a cost matrix with \( n + m \) column and rows where the \((i, j)\) entry is \( \|x_i - y_j\|^p \). When either \( x_i \) or \( y_j \) is a copy of a diagonal then this is the perpendicular distance. Each transportation plan corresponds to an assignment of rows to columns - a bijection between the points in \( X \) and those in \( Y \). The total cost of an assignment (or in other words bijection) \( \phi \) of tasks to people is \( \sum_{x \in X} \|x - \phi(x)\|^p \). Munkres algorithm gives us a bijection \( \phi \) that minimizes this cost. This means it gives an optimal pairing between \( X \) and \( Y \).

### 2.2 The curvature of \((D, d_p)\)

In order to understand the space of persistence diagrams it is useful to analyze its curvature. Alexandrov spaces are geodesic spaces with curvature bounded. They come in two different forms; their curvature bounded from above (also known as \( \text{CAT}(k) \) spaces) and their curvature bounded from below. A \( \text{CAT}(k) \) space is a geodesic space whose curvature is bounded from above by \( k \). A bound on curvature is defined in terms of comparison triangles. Consider a geodesic space \((X, d)\). Take three points \( x, y, z \) such that \( d(x, y) + d(y, z) + d(z, x) \leq \sqrt{2\pi/k} \) if \( k > 0 \) and these give us a triangle \( \Delta(x, y, z) \). For each \( k \in \mathbb{R} \) there is a model space \( M_k \) with constant curvature \( k \). We can build a comparison triangle \( \Delta(\tilde{x}, \tilde{y}, \tilde{z}) \) in the model space \( M_k \) with sides of the same length as the sides of \( \Delta(x, y, z) \). The curvature of \( X \) is bounded from below (above) if, for every triangle \( \Delta(x, y, z) \) in \( X \), the distances between the points on \( \Delta(x, y, z) \) are less than or equal (greater than or equal) the corresponding points in \( \Delta(\tilde{x}', \tilde{y}', \tilde{z}') \). For more details see [8]. \( \text{CAT} \)-spaces, in particular \( \text{CAT}(0) \) spaces, have nice properties. We first confirm that \((D, d_p)\) is not a \( \text{CAT}(k) \)-space.

**Proposition 2.** \((D, d_p)\) is not in \( \text{CAT}(k) \) for any \( k > 0 \) and any \( p \in [1, \infty) \).

**Proof.** If \((D, d_p) \in \text{CAT}(k)\) then there is a constant \( D_k \) such that for all \( X, Y \in (D, d_p) \) with \( d_p(X, Y)^2 < D_k \) there is a unique geodesic between them [8] between them. However, we can find \( X, Y \) arbitrarily close with two distinct geodesics. One example is taking \( X \) to be a diagram with two diagonally opposite corners of a square and \( Y \) a diagram with the other two corners. The horizontal and vertical paths are equally optimal and we may choose the square to be as small as we wish. \( \square \)

It was shown in [12] that \((D, d_2)\) is an Alexandrov space with curvature bounded below by zero. This is not the case for \( p \neq 2 \). From [11] we learn that a geodesic space \((X, d)\) is an Alexandrov spaces with curvature bounded from below by zero if, for any geodesic \( \gamma : [0, 1] \to X \) from \( X \) to \( Y \), and any \( Z \in X \) we have

\[
d(Z, \gamma(t))^2 \geq td(Z, Y)^2 + (1 - t)d(Z, X)^2 - t(1 - t)d(X, Y)^2.
\]

We will use different counterexamples for \( p \in [1, 2) \) and for \( p \in (2, \infty] \) to show that \((D, d_p)\) is not a non-negatively curved space whenever \( p \neq 2 \).
Let $p \in (2, \infty)$ and $t = 1/2$. Let $X, Y$ and $Z$ be a persistence diagram with only one off diagonal point each in them at $x = (1, 4), y = (1, 6)$ and $z = (0, 5)$ respectively. The midway point between $X$ and $Y$ (playing the role of $\gamma(1/2)$) is the diagram with the point $w = (1, 5)$. 

$$
\begin{align*}
&d_p(Z, \gamma(1/2))^p = \|z - w\|^p_p = 0^p + 1^p = 1 \\
d_p(Z, X)^p = \|z - x\|^p_p = 1^p + 1^p = 2 \\
d_p(Z, Y)^p = \|z - y\|^p_p = 1^p + 1^p = 2 \\
d_p(X, Y)^p = \|x - y\|^p_p = 0^p + 2^p = 2^p \\
\end{align*}
$$

$$
\frac{1}{2}d_p(Z, Y)^2 + \frac{1}{2}d_p(Z, X)^2 - \frac{1}{4}d_p(X, Y)^2 = 2^{2/p} - 1 < 1 = d_p(Z, \gamma(1/2))^2
$$

as $p > 2$. This contradicts equation (1) and hence $(D, d_p)$ is not an Alexandrov space with curvature bounded below by zero.

$$
\begin{align*}
&d_\infty(Z, \gamma(1/2)) = \|z - w\|_\infty = 1 \\
d_\infty(Z, X) = \|z - x\|_\infty = 1 \\
d_\infty(Z, Y) = \|z - y\|_\infty = 1 \\
d_\infty(X, Y) = \|x - y\|_\infty = 2
\end{align*}
$$

$$
\frac{1}{2}d_\infty(Z, Y)^2 + \frac{1}{2}d_\infty(Z, X)^2 - \frac{1}{4}d_\infty(X, Y)^2 = 0 < 1 = d_\infty(Z, \gamma(1/2))^2.
$$

This contradicts equation (1) and hence $(D, d_p)$ is not an Alexandrov space with curvature bounded below by zero.

Let $p \in [1, 2)$ and $t = 1/2$. Let $X, Y$ and $Z$ be a persistence diagram with only one off diagonal point each in them at $x = (0, 4), y = (2, 6)$ and $z = (0, 6)$ respectively. The midway point between $X$ and $Y$ (playing the role of $\gamma(1/2)$) is the diagram with the point $w = (1, 5)$. 

8
\[ d_p(Z, \gamma(1/2))^p = \|z - w\|_p^p = 1^p + 1^p = 2 \]
\[ d_p(Z, X)^p = \|z - x\|_p^p = 0^p + 2^p = 2^p \]
\[ d_p(Z, Y)^p = \|z - y\|_p^p = 2^p + 0^p = 2^p \]
\[ d_p(X, Y)^p = \|x - y\|_p^p = 2^p + 2^p = 2^{p+1} \]

\[
\frac{1}{2}d_p(Z, Y)^2 + \frac{1}{2}d_p(Z, X)^2 - \frac{1}{4}d_p(X, Y)^2 = 2^2 - 2^{2/p} < 2^{2/p} = d_p(Z, \gamma(1/2))^2
\]

as \( p < 2 \). This contradicts equation (1) and hence \((D, d_p)\) is not an Alexandrov space with curvature bounded below by zero.

### 3 The mean and the median as solutions of optimization

A statistic (singular) is a quantity that describes some attribute of a collection of data, summarizes some information about the data. It is found using some statistical algorithm with the set of data as input. More formally, statistical theory defines a statistic as a function of a sample where the function itself is independent of the sample’s distribution; that is, the function can be stated before realization of the data. The term statistic is used both for the function and for the value of the function on a given sample.

Basic descriptive statistics are often measures of central tendency and their corresponding measures of variability or dispersion. Measures of central tendency include the mean, median and mode. Measures of variability include the standard deviation, variance, the absolute deviation (total cost), average deviation, and the range of the values (distance between the minimum and maximum values of the variables).

Central tendencies (and their corresponding measures of variability) are solutions for optimizing different cost functions. These cost functions are based on \(p\)-Wasserstein metrics. We mainly care about when \( p = 1, 2 \) and \( \infty \). To motivate our cost functions for sets of persistence diagrams we will first recall common cost functions for sets of real numbers.
The mean of \(a_1, a_2, \ldots, a_N\) is the number \(\mu\) which minimizes the mean squared error

\[
F^\mathbb{R}_2(x) = \left( \frac{1}{N} \sum_{i=1}^{N} |a_i - x|^2 \right)^{1/2}
\]

The mean is thus

\[
\mu = \frac{1}{N} \sum_{i=1}^{N} a_i
\]

The standard deviation is the value \(F^\mathbb{R}_2(\mu)\).

The median of \(a_1, a_2, \ldots, a_N\), written in non-decreasing order, is the number \(m\) which minimizes the mean absolute deviation

\[
F^\mathbb{R}_1(x) = \frac{1}{N} \sum_{i=1}^{N} |a_i - x|.
\]

For \(N\) odd is unique and is \(a_{N+1}/2\). For \(N\) even can be any number in the interval \([a_N/a_{N+2}/2]\).

The range of \(a_1, a_2, \ldots, a_N\), written in non-decreasing order, is \(a_N - a_1\) and its midpoint is \(a_1 + a_N/2\). Consider the function

\[
F^\mathbb{R}_\infty(x) = \max_{i=1,\ldots,N} |a_i - x|
\]

which is the limit of \(F^\mathbb{R}_p(x)\) as \(p \to \infty\). The minimizer of \(F^\infty\) is the midpoint and its value is half the range. This represents the maximal cost of moving any point to the midpoint.

We wish to find the analogous cost functions and their corresponding central tendencies and measures of variability. To do this we need to use the appropriate metrics on \(D\) explored in section 2.1. Statistical qualities can be defined on the space of persistence diagrams by analogy using these different metrics. Given diagrams \(X_1, X_2, \ldots, X_N\) let

\[
F_p(Y) = \left( \frac{1}{N} \sum_{i=1}^{N} d_p(X_i, Y)^p \right)^{1/p} = \left( \frac{1}{N} \sum_{i=1}^{N} \inf_{\phi_i:Y \to X_i} \sum_{y \in Y} \|y - \phi_i(y)\|^p \right)^{1/p}
\]

and \(F^\infty(Y) = \sup_i d^\infty(Y, X_i)\).

The mean \(\mu\) is the diagram which minimizes \(F_2\) and \(F_2(\mu)\) is the standard deviation. The median \(m\) is the diagram which minimizes \(F_1\) and \(F_1(m)\) is the average deviation.

3.1 The mean and median of sets of points in the plane and copies of the diagonal

We want to gain some intuition over what the median and the mean looks like. To do this we will first restrict ourselves to understanding what the mean and median of sets of points in the plane and copies of the diagonal are.
Lemma 1. Let \((a_1, b_1), (a_2, b_2), \ldots, (a_k, b_k)\) be points in the plane. Let \(\hat{x}\) be the mean of \(a_1, a_2 \ldots a_k\) and \(\hat{y}\) be the mean of \(b_1, b_2, \ldots, b_k\). Then

\[
(x, y) := \left( \frac{k\hat{x} + (N - k)\frac{\hat{x} + \hat{y}}{2}}{N}, \frac{k\hat{y} + (N - k)\frac{\hat{x} + \hat{y}}{2}}{N} \right)
\]

is the unique point in \(\mathbb{R}^{2+}\) which minimizes

\[
\sum_{i=1}^{k} \| (x, y) - (a_i, b_i) \|^2_2 + \sum_{i=k+1}^{N} \| (x, y) - \Delta \|^2_2
\]

This proposition inspires the definition of a mean of a multiset of points in the plane and copies of the diagonal.

Definition 1. Let \(S\) be a multiset containing \((a_1, b_1), (a_2, b_2), \ldots, (a_k, b_k)\) and \(N - k\) copies of the diagonal and let \((x, y)\) be the point in \(\mathbb{R}^{2+}\) found in Lemma 1. We call this \((x, y)\) the mean of \(S\).

Proposition 3. Suppose \(k > N/2\). Let \((a_1, b_1), (a_2, b_2), \ldots, (a_k, b_k)\) be points in the plane. Let \((\tilde{x}, \tilde{y})\) be the point in \(\mathbb{R}^2\) where \(\tilde{x}\) is the median of \(a_1, a_2 \ldots a_k\) with \(N - k\) copies of \(-\infty\) and \(\tilde{y}\) is the median of \(b_1, b_2, \ldots, b_k\) with \(N - k\) copies of \(-\infty\). If \((\tilde{x}, \tilde{y})\) lies above the diagonal then \((\tilde{x}, \tilde{y})\) is the point in \(\mathbb{R}^{2+}\) which minimizes

\[
\sum_{i=1}^{k} \| (x, y) - (a_i, b_i) \|_1 + \sum_{i=k+1}^{N} \| (x, y) - \Delta \|_1.
\]

If \((\tilde{x}, \tilde{y})\) lies on or below the diagonal then \(f((x, y)) > \sum_{i=1}^{k} \| \Delta - (a_i, b_i) \|_1\) for all \((x, y) \in \mathbb{R}^{2+}\).

Proof. First observe that \(f\) is a convex function. This implies that to find the global minimum it is sufficient to find a local minimum.
Since $k > N/2$ we know that $\tilde{x}$ and $\tilde{y}$ are finite. Suppose that $(\tilde{x}, \tilde{y})$ lies above the diagonal. We want to show $(\tilde{x}, \tilde{y})$ is the minimum of $f$. Let $(u, v)$ be such that

$$|u| < \min_{a_i \neq \tilde{x}} |\tilde{x} - a_i|, \quad |v| < \min_{b_i \neq \tilde{y}} |\tilde{y} - b_i|,$$

and that $|u| + |v| < \| (\tilde{x}, \tilde{y}) - \Delta \|_1$. For such $(u, v)$ we have

$$\sum_{i=1}^k \| (\tilde{x} + u + \tilde{y} + v) - (a_i, b_i) \|_1 - \sum_{i=1}^k \| (\tilde{x}, \tilde{y}) - (a_i, b_i) \|_1$$

$$= |\{ i : a_i < \tilde{x} \} \cdot u + |\{ i : a_i > \tilde{x} \} \cdot (-u) + |\{ i : a_i = \tilde{x} \} \cdot |u| + (N - k)(-u)$$

$$+ |\{ i : b_i < \tilde{y} \} \cdot v + |\{ i : b_i > \tilde{y} \} \cdot (-v) + |\{ i : b_i = \tilde{y} \} \cdot |v| + (N - k)v$$

and

$$\| (\tilde{x} + u + \tilde{y} + v) - \Delta \|_1 - \| (\tilde{x}, \tilde{y}) - \Delta \|_1 = ((\tilde{y} + v) - (\tilde{x} + u)) - (\tilde{y} - \tilde{x}) = v - u$$

Together these imply that

$$f((\tilde{x} + u, \tilde{y} + v)) - f((\tilde{x}, \tilde{y}))$$

$$= |\{ i : a_i < \tilde{x} \} \cdot u + |\{ i : a_i > \tilde{x} \} \cdot (-u) + |\{ i : a_i = \tilde{x} \} \cdot |u| + (N - k)(-u)$$

$$+ |\{ i : b_i < \tilde{y} \} \cdot v + |\{ i : b_i > \tilde{y} \} \cdot (-v) + |\{ i : b_i = \tilde{y} \} \cdot |v| + (N - k)v$$

Since $\tilde{x}$ is the median of $a_1, a_2 \ldots a_k$ with $N - k$ copies of $\infty$ we know that

$$|\{ i : a_i > \tilde{x} \} | + (N - k)) - |\{ i : a_i < \tilde{x} \} | \leq \{ i : a_i = \tilde{x} \}$$

with a strict inequality when $N$ is odd. This implies that

$$|\{ i : a_i < \tilde{x} \} | \cdot u + |\{ i : a_i > \tilde{x} \} | \cdot (-u) + |\{ i : a_i = \tilde{x} \} | \cdot |u| + (N - k)(-u) \leq 0$$

again with a strict inequality when $N$ is odd and $u \neq 0$.

Similarly

$$|\{ i : b_i < \tilde{y} \} | \cdot v + |\{ i : b_i > \tilde{y} \} | \cdot (-v) + |\{ i : b_i = \tilde{y} \} | \cdot |v| + (N - k)v \geq 0$$

with a strict inequality when $N$ is odd and $v \neq 0$.

Thus $f((\tilde{x} + u, \tilde{y} + v)) \geq f((\tilde{x}, \tilde{y}))$ for $(\tilde{x} + u, \tilde{y} + v)$ sufficiently near $(\tilde{x}, \tilde{y})$. This implies that $(\tilde{x}, \tilde{y})$ is a local minimum and convexity implies that it must thus also be a global minimum of $f$ over the domain $\mathbb{R}^{2+}$ (we are not including the diagonal here and must be considered separately).

Now suppose that $(\tilde{x}, \tilde{y})$ lies on or below the diagonal. Let $(x, y) \in \mathbb{R}^{2+}$. Then either $x < \tilde{x}$ or $y > \tilde{y}$. Suppose that $x < \tilde{x}$. Let $x' \in (x, \tilde{x})$ with $(x', y) \in \mathbb{R}^{2+}$.

$$f((x, y)) - f((x', y)) = \sum_{i=1}^k \| (x, y) - (a_i, b_i) \|_1 - \| (x', y) - (a_i, b_i) \|_1 + \sum_{i=k+1}^N \| (x(t), y(t)) - \Delta \|_1$$

$$= \sum_{i=1}^k \| x - a_i \| - \| x' - a_i \| + \sum_{i=k+1}^N \| x' - x \|$$
Now \(|x - a_i| - |x' - a_i| = (x' - x)| whenever \(a_i \geq \tilde{x}\) and \((|x - a_i| - |x' - a_i|) \geq -(x' - x)\) for all \(i\). From \(\tilde{x}\) being the median of the \(a_i\) and \(N - k\) copies of \(\infty\) we know that

\[\{i : a_i \geq \tilde{x}\} + (N - k) > \{i : a_i < \tilde{x}\}.\]

Together we have

\[
\sum_{i=1}^{k}(|x - a_i| - |x' - a_i|) + \sum_{i=k+1}^{N}(x' - x) \geq \sum_{\{i:a_i \geq \tilde{x}\}}(x' - x) + \sum_{\{i:a_i \leq \tilde{x}\}}(x' - x) - (x' - x) \geq 0
\]

A similar argument shows that if \(y > \tilde{y}\) and \(y' \in (\tilde{y}, y)\) with \((x, y') \in \mathbb{R}^2^+\) then \(f((x, y)) > f((x, y'))\).

Thus we have \(f\) decreasing as we travel towards \((\tilde{x}, \tilde{y})\) while staying within \(\mathbb{R}^2^+\). Clearly if \((x, y)\) lies on the diagonal then

\[
\sum_{i=1}^{k}||(x, y) - (a_i, b_i)||_1 + \sum_{i=k+1}^{N}||(x, y) - \Delta||_1 > \sum_{i=1}^{k}||\Delta - (a_i, b_i)||_1
\]

with equality occurring if and only if \((a_i, b_i) = (x, y)\) for all \(i\) which is not allowed by our definition of a persistence diagram. \(\Box\)

**Lemma 2.** If \(k < N/2\) then

\[
\sum_{i=1}^{k}||(x, y) - (a_i, b_i)||_1 + \sum_{i=k+1}^{N}||(x, y) - \Delta||_1 > \sum_{i=1}^{k}||\Delta - (a_i, b_i)||_1
\]

for every point \((x, y) \in \mathbb{R}^2^+\). \(\Box\)

**Proof.** Since \(k < N/2\) we have \(\sum_{i=k+1}^{N}||(x, y) - \Delta||_1 > \sum_{i=1}^{k}||(x, y) - \Delta||_1\). By the triangle inequality we know \(||(x, y) - (a_i, b_i)||_1 + ||(x, y) - \Delta||_1 \geq ||\Delta - (a_i, b_i)||_1\). Together these imply the equation in the lemma. \(\Box\)

Using Proposition 2 and Lemma 2 we can formulate a useful definition of the median of a set containing points in the plane and copies of the diagonal.

**Definition 2.** Let \(S\) be the multiset of points \{\((a_1, b_1), (a_2, b_2), \ldots, (a_k, b_k)\)\} with \(N - k\) copies of the diagonal. Let \(\tilde{x}\) be the median of \(a_1, a_2, \ldots, a_k\) with \(N - k\) copies of \(\infty\) and let \(\tilde{y}\) be the median of \(b_1, b_2, \ldots, b_k\) with \(N - k\) copies of \(-\infty\).

- If \((\tilde{x}, \tilde{y})\) lies above the diagonal then we say the median of \(S\) is \((\tilde{x}, \tilde{y})\).
- If \((\tilde{x}, \tilde{y})\) lies on or below the diagonal (or is \((\infty, -\infty)\) which philosophically lies below the diagonal) then we say the median of \(S\) is the diagonal.
3.2 Characterizing the means and medians of sets of diagrams

Now that we understand what the mean or median of a set of points alongside copies of the diagonal are we can try to understand the mean or median of diagrams are. We need to worry about all the different choices of which point from each diagram get collected together in a manner similar finding the optimal bijections between pairs of diagrams.

Given a set of diagrams $X_1, \ldots, X_N$, a selection is a choice of one point from each diagram, where that point could be $\Delta$. A matching is a set of selections so that every off-diagonal point of every diagram is part of exactly one selection.

Our notation will be as follows. If $S$ is a selection then let $\mu_S$ be the mean of that selection as defined in definition 1 and let $m_S$ be the mean of that selection as defined in definition 2. A matching $G$ of $X_1, \ldots X_N$ is in fact a set of selections $G = \{S_j\}$. Let $\mu_G$ denote the persistence diagram which contains $\{\mu_{S_j} : S_j \in G\}$. Each matching $G$ thus produces a candidate $\mu_G$ for the mean and a candidate $m_G$ for the median. We will show that the mean and the median of $X_1, \ldots X_N$ must be some $\mu_G$ and $m_G$ respectively where $G$ and $G'$ are some matchings of $X_1, \ldots X_N$.

We found in [12] a characterization of the local minimums of $F_2$ when the observations are finitely many persistence diagrams each with only finitely many off-diagonal points.

Theorem 1. Let $X_1, \ldots, X_m$ be persistence diagrams with only finitely many off-diagonal points. $W = \{w_j\}$ is a local minimum of $F_2(Y) = \left(\frac{1}{m} \sum_{i=1}^m d_2(X_i, Y)^2\right)^{1/2}$ if and only if there is a unique optimal pairing from $W$ to each of the $X_i$, which we denote as $\phi_i$, and each $w_j$ is the mean of $\{\phi_i(w_j)\}_{i=1,2,\ldots,m}$.

We believe that a similar result may hold for local minimums of $F_1$. We do have a proof of a necessary condition.

Theorem 2. Let $Y \in \mathcal{D}$. For each $i$ let $\phi_i : Y \rightarrow X_i$ be an optimal bijection between $Y$ and $X_i$. For each $y \in Y$ we have a selection $\{\phi_i(y)\}$ (to make this well defined we think of the copies of the diagonal when $\phi_i^{-1}(x_j) = \Delta$ to each be disjoint). Let $G$ be the matching $\{\{\phi_i(y)\} : y \in Y\}$. If $Y$ is a local minimum of $F_1$ then $Y = m_G$.

Proof. Suppose that $Y$ is a local minimum of $F_1$ but that $Y \neq \mu_G$. 


\[ F_2(Y) = \frac{1}{N} \sum_i d_1(X_i, Y) \]
\[ = \frac{1}{N} \sum_i \sum_{y \in Y} \|y - \phi_i(y)\|_1 \]
\[ = \frac{1}{N} \sum_{y \in Y} \left( \sum_i \|y - \phi_i(y)\|_1 \right) \]

Let \( m_{\{\phi_i(y)\}} \) be the median of \( \{\phi_i(y)\} \). If \( Y \) is not \( m_G \) then \( y \neq m_{\{\phi_i(y)\}} \) for some \( y \in Y \). We need to split into cases depending on whether we are considering the diagonal or off-diagonal points.

If \( y = \Delta \) then \( \{\phi_i(y)\} \) contains at most one one diagonal point. By Lemma \( \ref{diagonal} \) we know that \( m_{\{\phi_i(y)\}} = \Delta \).

Suppose now that \( y \neq \Delta \). If \( \{\phi_i(y)\} \) is more that half copies the diagonal then by Lemma \( \ref{diagonal} \) we know \( m_{\{\phi_i(y)\}} = \Delta \). As we move \( z \) from \( y \) to the closest point on the diagonal \( \sum_{i: \phi(y)} \|z - \phi_i(y)\|_1 \) decreases less than \( \sum_{i: \phi(y)} \|z - \Delta\|_1 \) decreases and hence \( \sum_i \|z - \phi_i(y)\|_1 \) must be decreasing. This in turn implies that \( F_2 \) would also be decreasing as \( z \) moves towards the diagonal. Hence \( Y \) cannot be a local minimum.

Finally suppose that \( y \neq \Delta \) and more than half the points of \( \{\phi_i(y)\} \) are off the diagonal. Consider the point \((\tilde{x}, \tilde{y}) \in \mathbb{R}^2\) introduced in Proposition \( \ref{line} \). If \((\tilde{x}, \tilde{y})\) lies above the diagonal then by Proposition \( \ref{line} \) we know that \( \sum_i \|z - \phi_i(y)\|_1 \) decreases as \( z \) travels along a straight line from \( y \) to \( m_{\{\phi_i(y)\}} \). If \((\tilde{x}, \tilde{y})\) lies on or below the diagonal then the proof of Proposition \( \ref{line} \) shows that \( \sum_i \|z - \phi_i(y)\|_1 \) decreases as \( z \) moves from \( y \) to \( \Delta = m_{\{\phi_i(y)\}} \). In both cases this implies that \( F_1 \) would also decreasing as \( z \) moves from \( y \) towards \( m_{\{\phi_i(y)\}} \).

Thus by proof by contrapositive we have found our necessary condition for \( Y \) to be a local minimum.

Unlike in the situation of the mean we do not have the necessary condition of there being a unique optimal bijection from \( Y \) to \( X_i \) for each \( i \). This is because if we shift an observation \( a_i \) of a set \( a_1, \ldots, a_N \) of real numbers which is not central then we do not affect the median.

**Conjecture 1.** Let \( X_1, \ldots, X_m \) be persistence diagrams with only finitely many off-diagonal points. \( W = \{w_j\} \) is a local minimum of \( F_1(Y) = \frac{1}{m} \sum_{i=1}^m d_1(X_i, Y) \) if for any set of optimal pairings from \( W \) to each of the \( X_i \) which we denote as \( \phi_i \) and each \( w_j \) is the median of \( \{\phi_i(w_j)\}_{i=1,2,\ldots,m} \).

Theorems \( \ref{mean} \) and \( \ref{median} \) provide us with an (admittedly very slow) algorithm to find the mean and the median. We can consider the set of all matchings \( G \) and their candidates \( \mu_G \) and \( m_G \) for the mean and the median respectively. The mean is one of these \( \mu_G \) so we only need to compare the \( F_2(\mu_G) \) over all matchings \( G \). The median is one of these \( m_G \) so we only need to compare the \( F_1(m_G) \) over all matchings \( G \).

One qualitative difference between the mean and the median is the presence and absence of points with small persistence. Generally these points are heuristically thought of as noise. If we take the mean of a single point distance \( d \) from the diagonal and \( N - 1 \) copies of the
Lemma 3. Let $X_1, \ldots, X_N$ be persistence diagrams each with at most $K$ points in them. If $Y$ is a median of the $X_i$ then $Y$ has less than $2K$ points off the diagonal.

Proof. Let $y_1, y_2, \ldots, y_n$ be the off diagonal points in $Y$. Let $\phi_i$ be optimal bijections between $Y$ and the $X_i$. By Theorem we know that $y_j$ is the median of $\{\phi_i(y_j)\}$ for each $j$. By Lemma we know that for each $j$ the sets $\{\phi_i(y_j)\}$ must contain at least $(N+1)/2$ off diagonal points. This implies that $\bigcup_j \{\phi_i(y_j)\}$ must contain at least $(N + 1)n/2$ points.

Since the combined of total of all the off diagonal points in the $X_i$ is $NK$ we can conclude that $(N + 1)n/2 \leq NK$ and hence $n < 2K$.

In comparison it is possible for the mean of $N$ diagrams each with $K$ points to contain $NK$ off diagonal points.

4 Discontinuities of the mean and the median

Two unfortunate characteristics of both the mean and the median is that they are neither unique nor continuous. One way both the mean and median can fail to be unique and continuous comes down to the idea that which matching $G$ provides us with the optimal candidate for the mean or the median changes. This is illustrated in the Figures 7, 8, 9 and 10. We have three diagrams one of which is just the diagonal and the other are the blue and red. In this example as $z$ increases in the blue diagram travels across the optimal matching changes from $\{x_1, (1, z), \Delta\}$ and $\{x_2, \Delta, \Delta\}$ to $\{x_1, \Delta, \Delta\}$ and $\{x_2, (1, z), \Delta\}$ leading to a discontinuity to both the mean and the median (note it is in different locations that the switch occurs for the mean and the median). At the time it switches both matchings are equally optimal and hence we have non-uniqueness.
In Figure 7 $F_1$ (black) = $\frac{1}{3}((1 + (z - 2) + 1) + (2 + 0 + 0)) = (z + 2)/3$ and in Figure 8 $F_1$ (black) = $\frac{1}{3}((2 + 0 + 0) + (2 + (5 - z) + (z - 3))) = 2$. The median of the red, blue and purple (empty) diagrams is not continuous. When $z < 4$ the optimal matching is $\{(0, 2), (1, z), \Delta\}$ and $\{(3, 5), \Delta, \Delta\}$ (the matching used in Figure 7). When $z > 4$ then the optimal matching is $\{(0, 2), \Delta, \Delta\}$ and $\{(3, 5), (1, z), \Delta\}$ (the matching used in Figure 8). Both are optimal when $z = 4$ and as a result we do not have a unique median.

In Figure 9 $F_2$ (black) = $\frac{8639 - 3995z + 1268z^2}{6534}$ and in Figure 10 $F_2$ (black) = $\frac{191 - 58z + 7z^2}{36}$. The mean of the red, blue and purple (empty) diagrams is not continuous. $\frac{8639 - 3995z + 1268z^2}{6534} = \frac{191 - 58z + 7z^2}{36}$ at approximately $z = 3.99071$. When $z < 3.99071$ the optimal matching is $\{(0, 2), (1, z), \Delta\}$ and $\{(3, 5), \Delta, \Delta\}$ (the matching used in Figure 9). When $z > 3.99071$ then the optimal matching is $\{(0, 2), \Delta, \Delta\}$ and $\{(3, 5), (1, z), \Delta\}$ (the matching used in Figure 10). Both are optimal when $z = 3.99071$ and as a result we do not have a unique mean.
There is another way that the median can fail to be unique never happens with the mean. The mean is generically unique but the median is not. To show this rigorously we shall restrict ourselves to the case where we have $N$ diagrams each with only finitely many off diagonal points. Let $k_1, k_2, \ldots, k_N$ be non-negative integers. Let $U(k_1, k_2, \ldots, k_N)$ denote the space of sets of diagrams $X = \{X_1, X_2, \ldots, X_N\}$ such that $X_i$ has $k_i$ off diagonal points. $U(k_1, k_2, \ldots, k_N)$ is the quotient of $(\mathbb{R}^+)^{k_1+k_2+\ldots+k_N}$ by a finite group of symmetries $\Gamma$. There is a quotient map

$$q : (\mathbb{R}^+)^{k_1+k_2+\ldots+k_N} \to U(k_1, k_2, \ldots, k_N) = (\mathbb{R}^+)^{k_1+k_2+\ldots+k_N}/\Gamma.$$ 

Let $\lambda$ be Lebesgue measure on $(\mathbb{R}^+)^{k_1+k_2+\ldots+k_N}$ and let $\rho = q_*(\lambda)$ be the push forward of Lebesgue measure onto $U(k_1, k_2, \ldots, k_N)$. We will show that the measure of sets of diagrams in $U(k_1, k_2, \ldots, k_N)$ which do not have a unique mean is zero. In comparison we can show that the measure of the sets of diagrams in $U(k_1, k_2, \ldots, k_N)$ which do not have a unique median has positive measure.

**Proposition 4.** The sets of diagrams in $U(k_1, k_2, \ldots, k_N)$ which do not have a unique mean has measure zero.

**Proof.** Let $\hat{A}$ be the set of sets of diagrams in $U(k_1, k_2, \ldots, k_N)$ which do not have a unique mean. Now $\rho(\hat{A}) = \lambda(q^{-1}(\hat{A}))$ so showing $\hat{A}$ has measure zero it is equivalent to showing $\lambda(q^{-1}(\hat{A})) = 0$. Let $A = q^{-1}(\hat{A})$. $A$ is the set of vectors of labelled diagrams which do not have a unique mean.

By vectors of labelled diagrams we mean objects of the form $(X_1, X_2, \ldots, X_N)$ where we label the off diagonal points within each diagram, $x_i \in X_i$. We want to show $\lambda(A) = 0$.

Let $S$ be a selection containing the points $\{(a_1, b_1), (a_2, b_2), \ldots, (a_k, b_k)\}$ with $N-k$ copies of the diagonal. Recall from the definition that

$$\mu_S = \left(\frac{k\hat{x} + (N-k)\hat{y}}{N}, \frac{k\hat{y} + (N-k)\hat{x}}{N}\right)$$

where $\hat{x}$ and $\hat{y}$ are the means of $a_1, a_2, \ldots, a_k$ and $b_1, b_2, \ldots, b_k$ respectively. For each selection $S$ let $f_S(X) = \sum_{x \in S} \|x - \mu_S\|^2$. Then $f_S(X)$ is a quadratic function of $a_1, a_2, \ldots, a_k, b_1, b_2, \ldots, b_k$. If $X$ has more than one mean then by Theorem [1] there must be matchings $G_1, G_2$ such that $m_{G_1} \neq m_{G_2}$ are both means and

$$\sum_{S \in G_1} f_S(X) = F_2(m_{G_1})^2 = F_2(m_{G_2})^2 = \sum_{S \in G_2} f_S(X). \quad (2)$$

For each pair of matchings $G_1, G_2$ let

$$A(G_1, G_2) = \left\{ X = (X_1, X_2, \ldots, X_N) : \sum_{S \in G_1} f_S(X) = \sum_{S \in G_2} f_S(X) \right\}.$$  

From (2) we have

$$A \subseteq \bigcup_{G_1, G_2 \text{ matchings}} A(G_1, G_2).$$
Since the points in \( A(G_1, G_2) \) satisfy a single equation which is quadratic in each of the coordinates we know that either \( A(G_1, G_2) = (\mathbb{R}^2)^{k_1+k_2+\ldots+k_N} \) or that \( \lambda(A(G_1, G_2)) = 0 \). It is clear that there exists a vector of labelled persistence diagrams \( X = (X_1, X_2, \ldots X_N) \in (\mathbb{R}^2)^{k_1+k_2+\ldots+k_N} \) such that \( X \notin A(G_1, G_2) \). Thus we can conclude that \( A(G_1, G_2) \neq (\mathbb{R}^2)^{k_1+k_2+\ldots+k_N} \) and hence \( \lambda(A(G_1, G_2)) = 0 \).

There are only finitely many matchings so \( \lambda(A(G_1, G_2)) = 0 \) for all pairs of matchings \( G_1, G_2 \) implies that \( \lambda(A) = 0 \). \( \square \)

This proof of generic uniqueness contrasts sharply to the case of the median which is not generically unique.

**Proposition 5.** Let \( k_1, k_2, \ldots, k_{(N+1)/2} \geq 2 \). The sets of diagrams in \( U(k_1, k_2, \ldots k_N) \) which do not have a unique median has positive measure.

**Proof.** We will first illustrate this with the case \( U(2, 2, 0) \). This example shows the idea of the general case. Suppose \( X_1 \) have two off diagonal points \( (a_1, a_2) \) and \( (a_2, b_2) \), \( X_2 \) has two off diagonal points \( (c_1, d_1) \) and \( (c_2, d_2) \), and \( X_3 \) has no off diagonal points. Further suppose that

\[
\begin{align*}
& a_1, a_2 < c_1, c_2 \leq b_1, b_2 < d_1, d_2. 
\end{align*}
\]

The possible matchings are

\[
\begin{align*}
G_1 & = \{(a_1, b_1), (c_2, d_2), \Delta\}, \{(a_2, b_2), (c_1, d_1), \Delta\} \\
G_2 & = \{(a_1, b_1), (c_1, d_1), \Delta\}, \{(a_2, b_2), (c_2, d_2), \Delta\} \\
G_3 & = \{(a_1, b_1), \Delta, \Delta\}, \{(\Delta, c_2, d_2), \Delta\}, \{(a_2, b_2), (c_1, d_1), \Delta\} \\
G_4 & = \{(a_1, b_1), (c_2, d_2), \Delta\}, \{(a_2, b_2), \Delta, \Delta\}, \{(\Delta, c_1, d_1), \Delta\} \\
G_5 & = \{(a_2, b_2), \Delta, \Delta\}, \{(\Delta, c_2, d_2), \Delta\}, \{(a_1, b_1), (c_1, d_1), \Delta\} \\
G_6 & = \{(a_2, b_2), (c_2, d_2), \Delta\}, \{(a_1, b_1), \Delta, \Delta\}, \{(\Delta, c_1, d_1), \Delta\} \\
G_7 & = \{(a_1, b_1), \Delta, \Delta\}, \{(\Delta, c_2, d_2), \Delta\}, \{(a_2, b_2), \Delta, \Delta\}, \{(\Delta, c_1, d_1), \Delta\} \\
\end{align*}
\]

From Figure 11 and Figure 12 we can see that

\[
\begin{align*}
F_1(m_{G_1}) & = (c_2 - a_1) + (d_2 - b_1) + (b_1 - c_2) + (c_1 - a_2) + (d_1 - b_2) + (b_2 - c_1) \\
& = -a_1 + d_2 - a_2 + d_1 \\
F_1(m_{G_2}) & = (c_1 - a_1) + (d_1 - b_1) + (b_1 - c_1) + (c_2 - a_2) + (d_2 - b_2) + (b_2 - c_2) \\
& = -a_1 + d_1 - a_2 + d_2 \\
& = F_1(m_{G_1}) \\
\end{align*}
\]

We can show that \( F_1(m_{G_k}) \geq -a_1 + d_1 - a_2 + d_2 \) for \( k = 3, 4, 5, 6, 7 \). For example, \( G_3 \) contains one off diagonal point located at \( (c_1, b_2) \) and

\[
\begin{align*}
F_1(m_{G_3}) & = \| (a_2, b_2) - (c_1, b_2) \|_1 + \| (c_1, d_1) - (c_1, b_2) \|_1 + \| \Delta - (c_1, b_2) \|_1 \\
& \quad + \| (a_1, b_1) - \Delta \|_1 + \| (c_2, d_2) - \Delta \|_1 \\
& = (c_1 - a_2) + (d_1 - b_2) + (b_2 - c_1) + (b_1 - a_1) + (d_2 - c_2) \\
& = (-a_1 + d_1 - a_2 + d_2) + b_1 - c_2 \\
& \geq -a_1 + d_1 - a_2 + d_2. \\
\end{align*}
\]
Figure 11: \( m_{(1,2)} := (c_2, b_1) \) is the mean of the selection \( S_{(1,2)} := \{(a_1, b_1), (c_2, d_2), \Delta\} \) and \( m_{(2,1)} := (c_1, b_2) \) is the mean of the selection \( S_{(2,1)} := \{(a_2, b_2), (c_1, d_1), \Delta\} \). This implies that the black diagram is \( m_{G_1} \) where \( G_1 = \{S_{(1,2)}, S_{(1,2)}\} \).

We have \( F_1(m_{G_1}) = -a_1 + d_2 - a_2 + d_1 \).

Figure 12: \( m_{(1,1)} := (c_1, b_1) \) is the mean of the selection \( S_{(1,1)} := \{(a_1, b_1), (c_1, d_1), \Delta\} \) and \( m_{(2,2)} := (c_2, b_2) \) is the mean of the selection \( S_{(2,2)} := \{(a_2, b_2), (c_2, d_2), \Delta\} \). This implies that the black diagram is \( m_{G_2} \) where \( G_2 = \{S_{(1,1)}, S_{(2,2)}\} \).

We have \( F_1(m_{G_2}) = -a_1 + d_1 - a_2 + d_2 \).
The last inequality is because $b_1 \geq c_2$ by assumption. The calculations for $k = 4, 5, 6, 7$ are similar.

This implies that $m_{G_1}$ and $m_{G_2}$ are both means of X. If $b_1 \neq b_2$ or $c_1 \neq c_2$ these means are distinct and thus we do not have a unique mean. The measure of such sets of diagrams $\{X_1, X_2, X_3\}$ has non-zero measure in $U(2, 2, 0)$.

We will now sketch an extension of this example to the case where $k_1, k_2, \ldots, k_{(N+1)/2} > 2$. This is illustrated in Figure 13. Put $(a_1, b_1)$ and $(a_2, b_2)$ from $X_1$ in the red region. Put $(c_1, d_1)$ and $(c_2, d_2)$ from $X_2$ in the blue region. Put exactly two points from each of $X_3, X_4 \ldots X_{(N+1)/2}$ in the purple region. Put every other off diagonal point in the orange region. If $m$ is a median then 2 of its off diagonal points will be $\{(c_1, b_1), (c_2, b_2)\}$ or $\{(c_2, b_1), (c_1, b_2)\}$. Another median $\tilde{m}$ is the same as $m$ but switching $\{(c_1, b_1), (c_2, b_2)\}$ for $\{(c_2, b_1), (c_1, b_2)\}$ or vice versa.

Let $m$ be a mean of $\{X_1, X_2, \ldots X_N\}$. We will show that there is a matching $G$ such that every selection $S \in G$ cannot contain both points in the orange region as well as points in the combined red, blue and purple regions such that $m = m_G$. Certainly $m = m_G$ for some matching. Suppose that there is some selection $S \in G$ contains $l \neq 0$ off diagonal points in the combined red, blue and purple regions as well $q \neq 0$ off diagonal points in the orange region and that $m_S$ is not the diagonal.

If both the $x$-coordinate of $m_S$ is determined by a point in orange region then any optimal matching between $m_G$ and any of the $X_i$ would necessarily send $m_S$ to either a point in the orange region or to the diagonal. This contradicts the result in Theorem 2 that $m_S$ is the median of $\{\phi_i(m_S)\}$. Similarly if the $y$-coordinate of $m_S$ is determined by some point in the red, blue or purple regions then then any optimal matching between $m_G$ and any of the $X_i$ would necessarily send $m_S$ to either a point in the red, blue or purple regions or to the diagonal. This again contradicts the result in Theorem 2 that $m_S$ is the median of $\{\phi_i(m_S)\}$. Thus we know that the $x$-coordinate must be determined by a point in the red, blue purple
regions and hence \( l \geq (N + 1)/2 \). Simultaneously we know that the \( y \)-coordinate must be determined by a point in the orange region and hence \( q \geq (N + 1)/2 \). However this would imply that \( l + q > N \) which is impossible. We can conclude that if \( S \) is a selection of \( G \) where \( m_G \) is a mean and \( S \) contains both points in the orange region as well as points in the combined red, blue and purple regions then \( m_S \) is the diagonal. We can replace \( S = \{s_1, s_2, \ldots, s_N\} \) with multiple selections - each containing only one off diagonal point \( s_i \).

We can split our diagrams \( X_i \) into \( Y_i \) and \( Z_i \) where \( Y_i \) has the points in \( X_i \) that lie in the red, blue or purple regions and \( Z_i \) has the points in \( X_i \) that lie in the orange region. We have shown that any mean \( m \) of the \( X_i \) is the amalgamation of a mean of the \( Y_i \) and a mean of the \( Z_i \) (where by amalgamation of persistence diagrams \( A \) and \( B \) we mean the set of off diagonal points is the union of the sets of the off diagonal points in \( A \) and \( B \)). We can ignore the points in the orange region from now on. Non-unique medians of the \( Y_i \) will imply non-unique medians of the \( X_i \).

The proof that the \( Y_i \) has non-unique medians is very similar to the example seen in Figure 11 and Figure 12. The selections involve one splitting the points in the purple region from \( Y_3, \ldots, Y_{(N-3)/2} \) into two distinct sets each set containing one point from each of the \( Y_i \). Interchanging the the two points within these \( Y_i \) \((i > 3)\) will not affect the location of \( m_S \) as their \( x \)-coordinates are always less than those in the red and blue regions and their \( y \)-coordinates are always greater than those in the red and blue regions. We then have a choice that will lead to two different means. If we add \( \{(a_1, b_1), (c_2, d_2)\} \) to one selection and \( \{(a_2, b_2), (c_1, d_1)\} \) to the other we get \( m_{G_1} = \{(c_1, b_1), (c_1, b_2)\} \). Alternatively if we add \( \{(a_1, b_1), (c_1, d_1)\} \) to one selection and \( \{(a_2, b_2), (c_2, d_2)\} \) to the other we get \( m_{G_2} = \{(c_1, b_1), (c_2, b_2)\} \). That \( m_{G_1} \) and \( m_{G_2} \) both minimize \( F_1 \) follows the same (but longer) calculations as the case of \( U(2, 2, 0) \). Now \( m_{G_1} \) and \( m_{G_2} \) are distinct whenever \( a_1 \neq a_2 \) or \( c_1 \neq c_2 \). The set of \( Y_1, Y_2, \ldots, Y_N \) where \( a_1 = a_2 \) and \( c_1 = c_2 \) is of zero measure and so our non-uniqueness result holds on a set of positive measure.

5 Discussion and further directions

This paper finds the natural definition of the mean and the median of a set of diagrams. This is through considering the cost functions analogous to those of samples of real numbers and defining these central tendencies to be the solutions for optimizing these cost functions. We then characterize what the local minimums of these different cost functions and in doing so characterize the mean and the median. Many parallels are shown between the mean and the median. This suggests that some future directions could involve extending work that has been done on the mean to the corresponding results for the median. For example, the discontinuity and lack of uniqueness of both the mean and the median is unfortunate. It makes statistical inference much harder. One possible workaround is to consider some other definition of the mean and the median. In [10] they explore an alternative probabilistic definition of the mean which combines the tradition (Fréchet) mean used in this paper with the notion of a shaking hand equilibrium in game theory. We feel that a similar idea would work to create a probabilistic definition of the median.

The space of persistence diagrams is of interest for its own sake. We have proved some results about the curvature and structure of this space. It would be interesting to see if \((\mathcal{D}, d_p)\), for \( p \neq 2 \) does have some bound on the curvature from below. It is not bounded from
below by zero but it may be by something else.

References

[1] Chanderjit Bajaj. The algebraic degree of geometric optimization problems. *Discrete & Computational Geometry*, 3(1):177–191, 1988.

[2] Sivaraman Balakrishnan, Brittany Fasy, Fabrizio Lecci, Alessandro Rinaldo, Aarti Singh, and Larry Wasserman. Statistical inference for persistent homology. *arXiv preprint arXiv:1303.7117*, 2013.

[3] Andrew J Blumberg, Itamar Gal, Michael A Mandell, and Matthew Pancia. Persistent homology for metric measure spaces, and robust statistics for hypothesis testing and confidence intervals. *arXiv preprint arXiv:1206.4581*, 2012.

[4] Peter Bubenik. Statistical topology using persistence landscapes. *arXiv preprint arXiv:1207.6437*, 2012.

[5] Peter Bubenik and Peter T Kim. A statistical approach to persistent homology. *Homology, Homotopy and Applications*, 9(2):337–362, 2007.

[6] Frédéric Chazal, Marc Glisse, Catherine Labruère, and Bertrand Michel. Optimal rates of convergence for persistence diagrams in topological data analysis. *arXiv:1305.6239*.

[7] Robert Ghrist. Barcodes: the persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1):61–75, 2008.

[8] William A Kirk. Geodesic geometry and fixed point theory. In *Seminar of Mathematical Analysis (Malaga/Seville, 2002/2003)*, volume 64, pages 195–225, 2003.

[9] Yuriy Mileyko, Sayan Mukherjee, and John Harer. Probability measures on the space of persistence diagrams. *Inverse Problems*, 27(12):124007, 2011.

[10] Elizabeth Munch, Paul Bendich, Katharine Turner, Sayan Mukherjee, Jonathan Mattingly, and John Harer. Probabilistic Fréchet means and statistics on vineyards. *arXiv preprint arXiv:1307.6530*, 2013.

[11] Shin-ichi Ohta. Barycenters in alexandrov spaces of curvature bounded below. 2012.

[12] Katharine Turner, Yuriy Mileyko, Sayan Mukherjee, and John Harer. Fréchet means for distributions of persistence diagrams. *arXiv preprint arXiv:1206.2790*, 2012.

[13] Katharine Turner and Andrew Robinson. Null hypothesis testing in the space of persistence diagrams. *in preparation*. 

23