Statistics of Persistence Diagrams

Katharine Turner

University of Chicago

kate@math.uchicago.edu

What is a statistic?

A statistic (singular) is a quantity that describes some attribute of a sample of data. It summarizes some information about the data.

What is a statistic?

A statistic (singular) is a quantity that describes some attribute of a sample of data. It summarizes some information about the data.

A statistic is found using some statistical algorithm with the set of data as input.

What is a statistic?

A statistic (singular) is a quantity that describes some attribute of a sample of data. It summarizes some information about the data.

A statistic is found using some statistical algorithm with the set of data as input.

Basic descriptive statistics are often measures of central tendency and their corresponding measures of variability or dispersion.

What is a statistic?

A statistic (singular) is a quantity that describes some attribute of a sample of data. It summarizes some information about the data.

A statistic is found using some statistical algorithm with the set of data as input.

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.

What is a statistic?

A statistic (singular) is a quantity that describes some attribute of a sample of data. It summarizes some information about the data.

A statistic is found using some statistical algorithm with the set of data as input.

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* (or variance), the *absolute deviation*, and the range of the values (distance between the minimum and maximum values of the variables).

Central tendencies as function minimisers

Central tendencies (and their corresponding measures of variability) are solutions for optimizing different cost functions. These cost functions are based on L^p norms on function spaces. We mainly care about when $p = 1, 2$ and ∞.

Central tendencies as function minimisers

Central tendencies (and their corresponding measures of variability) are solutions for optimizing different cost functions. These cost functions are based on L^p norms on function spaces. We mainly care about when $p = 1, 2$ and ∞.

Given a sample $a_1, a_2, \ldots, a_n$, some familiar statistical quantities are minimisers of functions of the form

$$F_p(x) = \left(\sum_{i=1}^{N} |a_i - x|^p \right)^{1/p} \qquad F_\infty(x) = \sup_i |a_i - x|$$

Central tendencies as function minimisers

Central tendencies (and their corresponding measures of variability) are solutions for optimizing different cost functions. These cost functions are based on L^p norms on function spaces. We mainly care about when $p = 1, 2$ and ∞.

Given a sample $a_1, a_2, \ldots, a_n$, some familiar statistical quantities are minimisers of functions of the form

$$F_p(x) = \left(\sum_{i=1}^{N} |a_i - x|^p \right)^{1/p} \qquad F_\infty(x) = \sup_i |a_i - x|$$

- $p = 2$: the mean minimizes the mean squared error

Central tendencies as function minimisers

Central tendencies (and their corresponding measures of variability) are solutions for optimizing different cost functions. These cost functions are based on L^p norms on function spaces. We mainly care about when $p = 1, 2$ and ∞.

Given a sample $a_1, a_2, \ldots, a_n$, some familiar statistical quantities are minimisers of functions of the form

$$F_p(x) = \left(\sum_{i=1}^{N} |a_i - x|^p \right)^{1/p} \qquad F_\infty(x) = \sup_i |a_i - x|$$

- $p = 2$: the mean minimizes the mean squared error
- $p = 1$: the median minimizes absolute deviation

Central tendencies as function minimisers

Central tendencies (and their corresponding measures of variability) are solutions for optimizing different cost functions. These cost functions are based on L^p norms on function spaces. We mainly care about when $p = 1, 2$ and ∞.

Given a sample $a_1, a_2, \ldots, a_n$, some familiar statistical quantities are minimisers of functions of the form

$$F_p(x) = \left(\sum_{i=1}^{N} |a_i - x|^p \right)^{1/p} \qquad F_\infty(x) = \sup_i |a_i - x|$$

- $p = 2$: the mean minimizes the mean squared error
- $p = 1$: the median minimizes absolute deviation
- $p = \infty$: the mid-range point minimizes the maximum absolute deviation

Mean and variance for data on the real line

The mean of $a_1, a_2, \ldots a_N$ is the number μ which minimizes the of the mean squared error

$$F_2(x) = \left(\sum_{i=1}^{N} |a_i - x|^2 \right)^{1/2}$$

The mean is thus

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

The standard deviation is the value $F_2(\mu)$.

Median and absolute deviation for data on the real line

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

$$F_1(x) = \sum_{i=1}^{N} |a_i - x|.$$

For N odd is unique and is $a_{\frac{N+1}{2}}$. For N even can be any number in the interval $[a_{\frac{N}{2}}, a_{\frac{N+2}{2}}]$. The total cost of moving everything from the sample data to the median is $F_1(m)$

Midrange point and range for data on the real line

The range of $a_1, a_2, \ldots, a_N$, written in non-decreasing order, is $a_N - a_1$ and its midpoint is $\frac{a_1+a_N}{2}$. Consider the function

$$F_\infty(x) = \max\{|a_i - x|\}.$$

The minimizer of F_∞ is the midpoint and its value is half the range. This represents the maximal cost of moving any point to the midpoint.

Persistence diagrams

Persistence diagrams describe how the topology changes with the progressive inclusion of sublevel sets.

Persistence diagrams

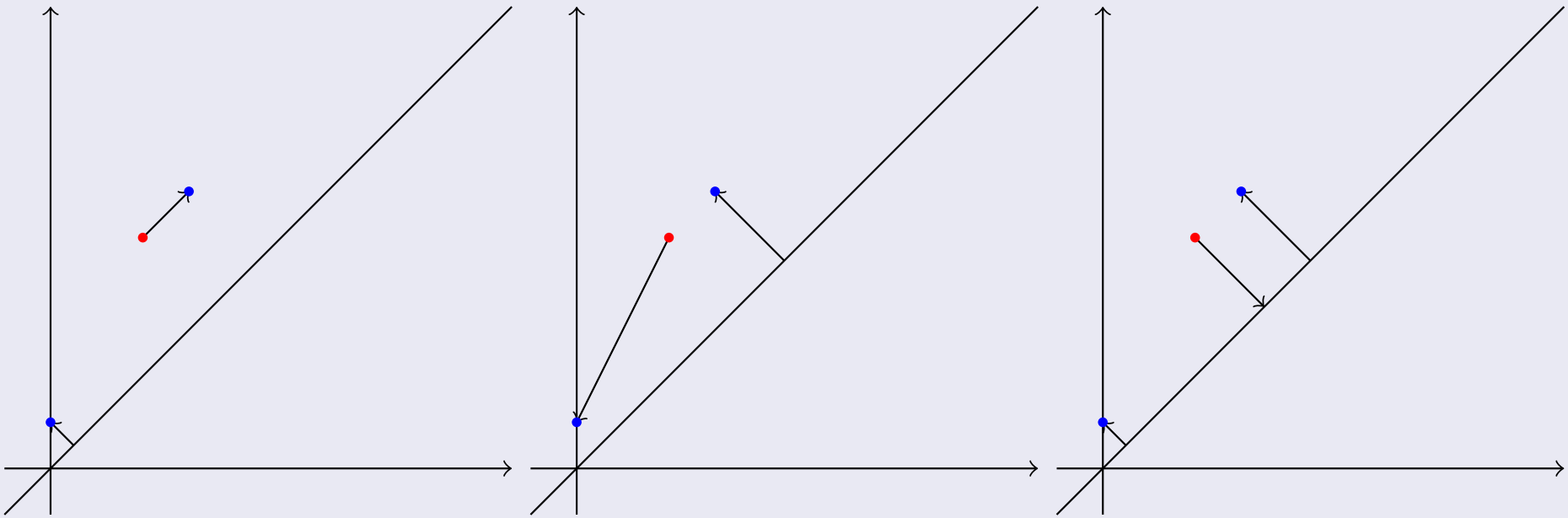
Persistence diagrams describe how the topology changes with the progressive inclusion of sublevel sets.

Definition

A persistence diagram is a countable multiset of points in $\overline{\mathbb{R}}^2$ along with the diagonal $\Delta = \{(x, y) \in \mathbb{R}^2 \mid x = y\}$, where each point on the diagonal has infinite multiplicity. We require some niceness assumptions.

Given two diagrams X, Y we can considers bijections ϕ between the points in X and the points in Y. Bijections always exist because there are countably many points at every location on the diagonal. We only need to consider bijection where off-diagonal points are either paired with off-diagonal points or with the point on the diagonal that is closest to it.

Example: $\phi : X(\text{red}) \to Y(\text{blue})$



Distance between diagrams

There are many choices of metrics in the space of persistence diagrams just like there are different choices of metric on spaces of functions. We will consider three choices which are analogous to L^1, L^2 and L^∞ on the space of real valued functions.
One family of distances are

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

for $p \geq 1$ and taking limits for $p = \infty$ getting

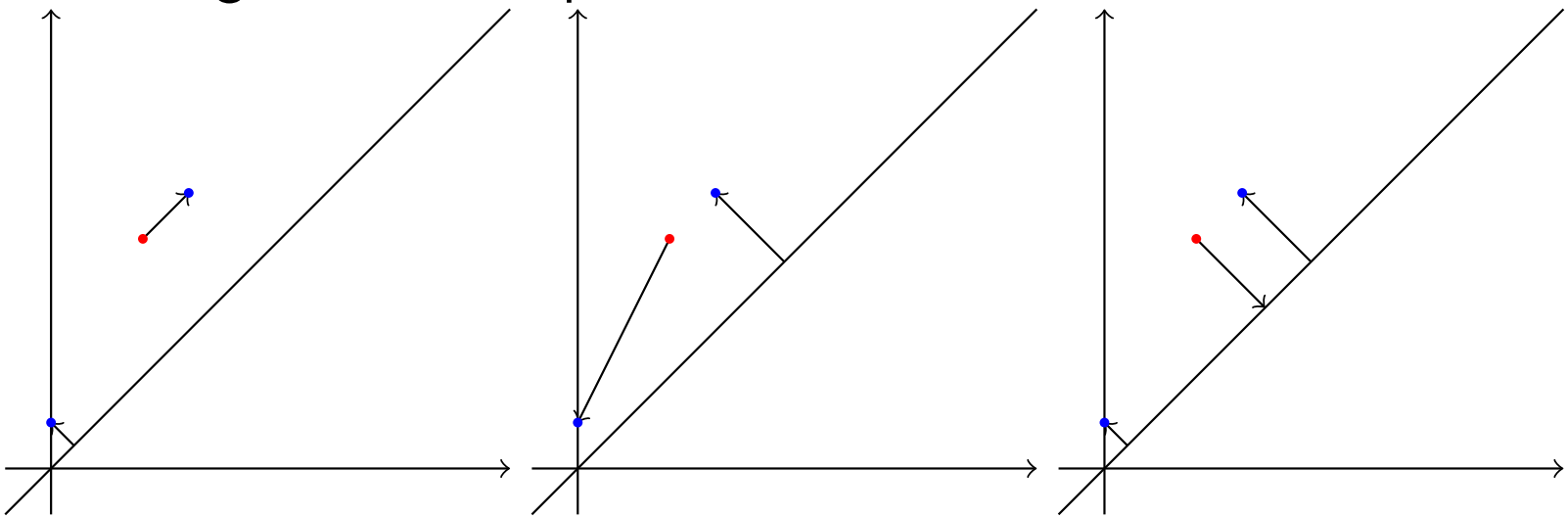
$$d_\infty(X,Y) = \inf_{\phi: X \to Y} \max\{\|x - \phi(x)\|_\infty\}$$

The optimal pairing

optimal pairing

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

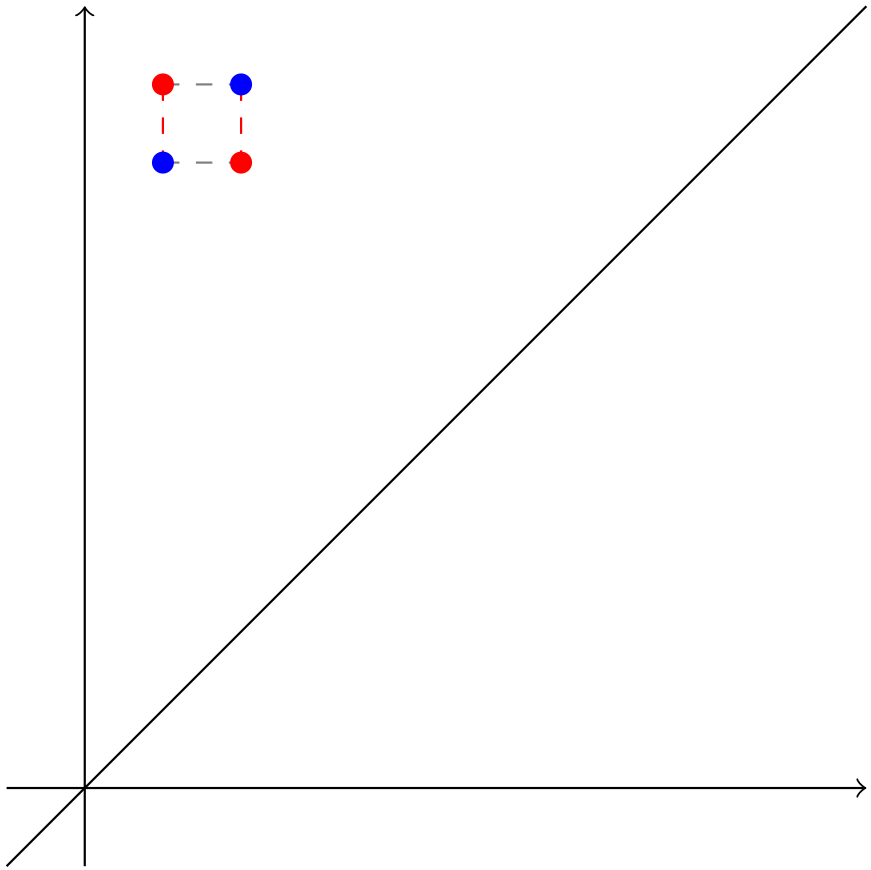
Returning to our example.



The optimal choice is the first one for all values of p.

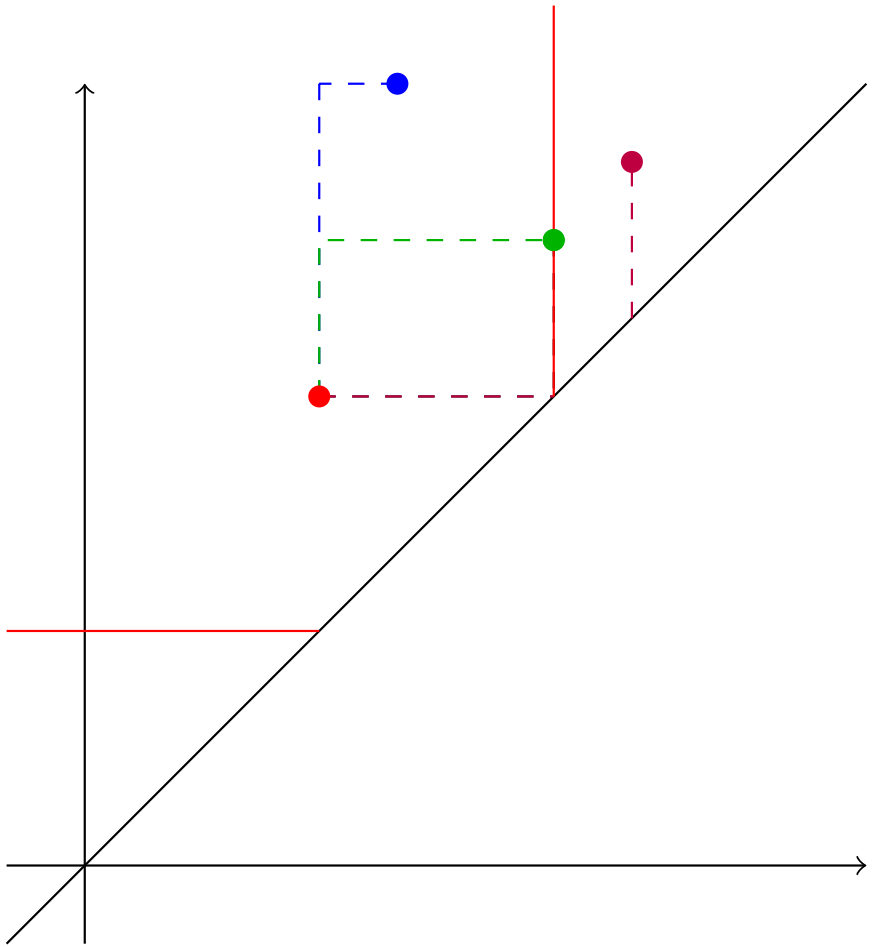
Non-uniqueness away from the diagonal

This example works for every $p \in [1, \infty]$. Matching the points vertically or horizontally involves the same cost.



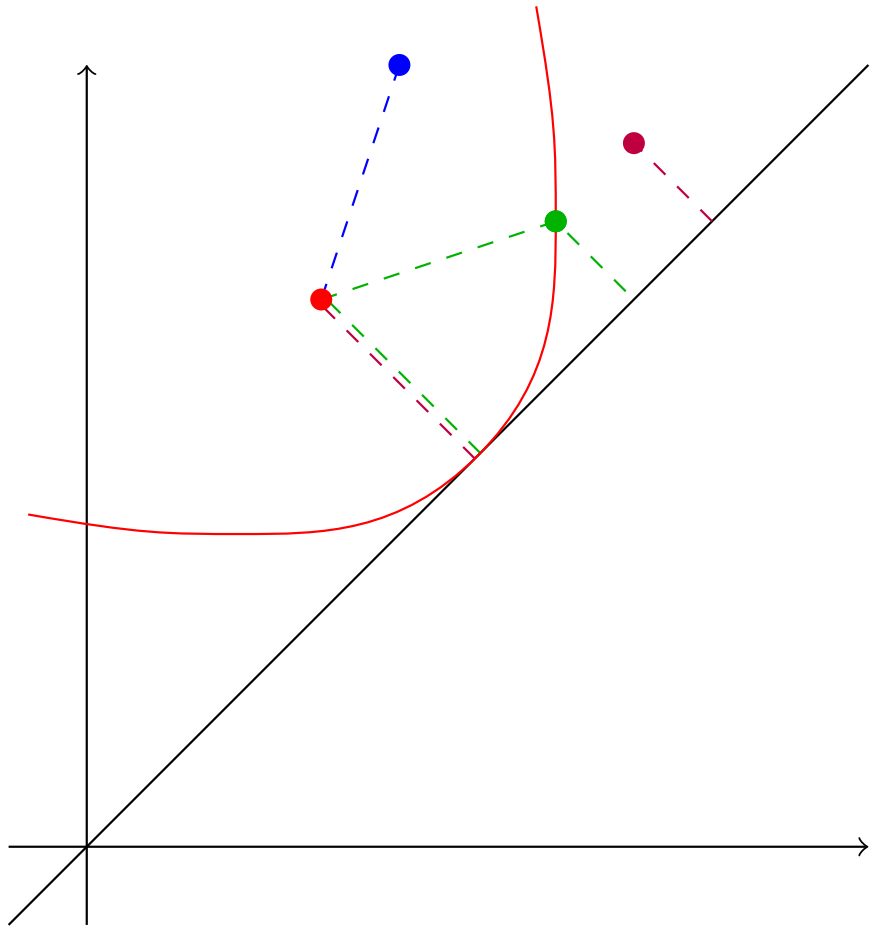
Non-uniqueness involving the diagonal, $p = 1$

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).



Non-uniqueness involving the diagonal, $p = 2$

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.



Statistical qualities can thus be defined on the space of persistence diagrams by analogy using these different distances. Given diagrams $X_1, X_2 \dots X_N$ let

$$F_p(Y) = \left(\sum_{i=1}^{N} d_p(X_i, Y)^p \right)^{1/p} \qquad F_\infty(Y) = \sup_i d_\infty(Y, X_i).$$

- The mean μ 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 absolute deviation.
- The "midpoint" $\hat{m}$ is the diagram which minimizes F_∞ and $F_\infty(\hat{m})$ is the maximal absolute deviation.

"Mean" of points in $\mathbb{R}^2$ and copies of the diagonal

Lemma

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 point in $\mathbb{R}^2$ which minimizes

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

"Mean" of points in $\mathbb{R}^2$ and copies of the diagonal

Lemma

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

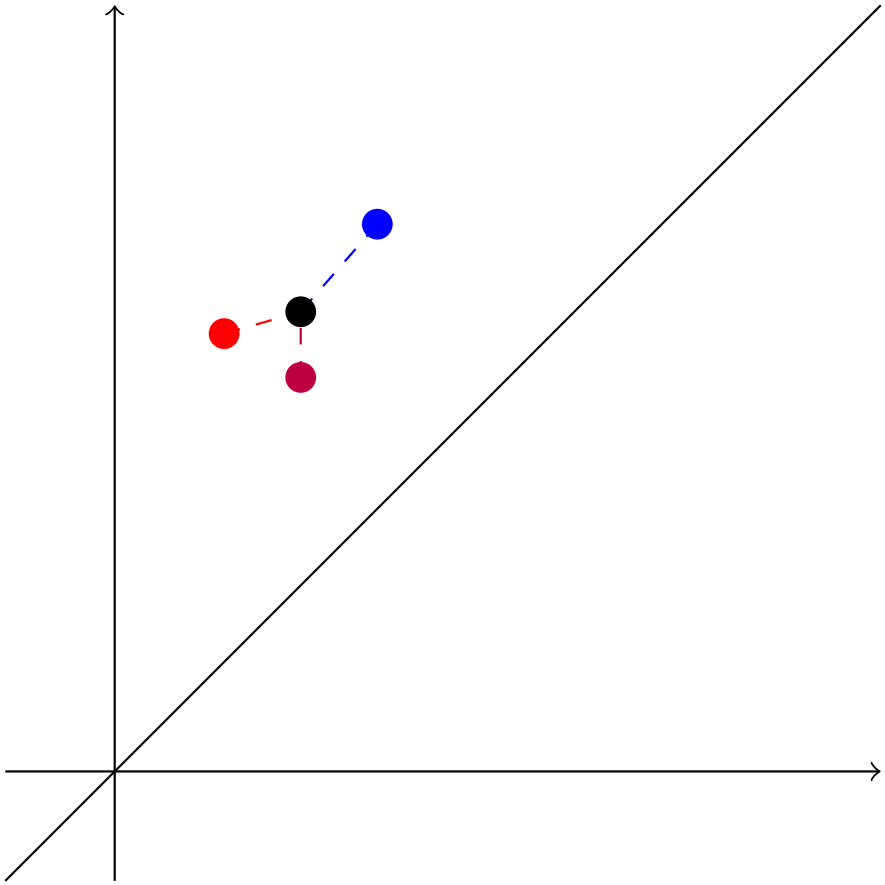
$$(x, y) := \left(\frac{k\hat{x} + (n-k)\frac{\hat{x}+\hat{y}}{2}}{n}, \frac{k\hat{y} + (n-k)\frac{\hat{x}+\hat{y}}{2}}{n} \right)$$

is the point in $\mathbb{R}^2$ which minimizes

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

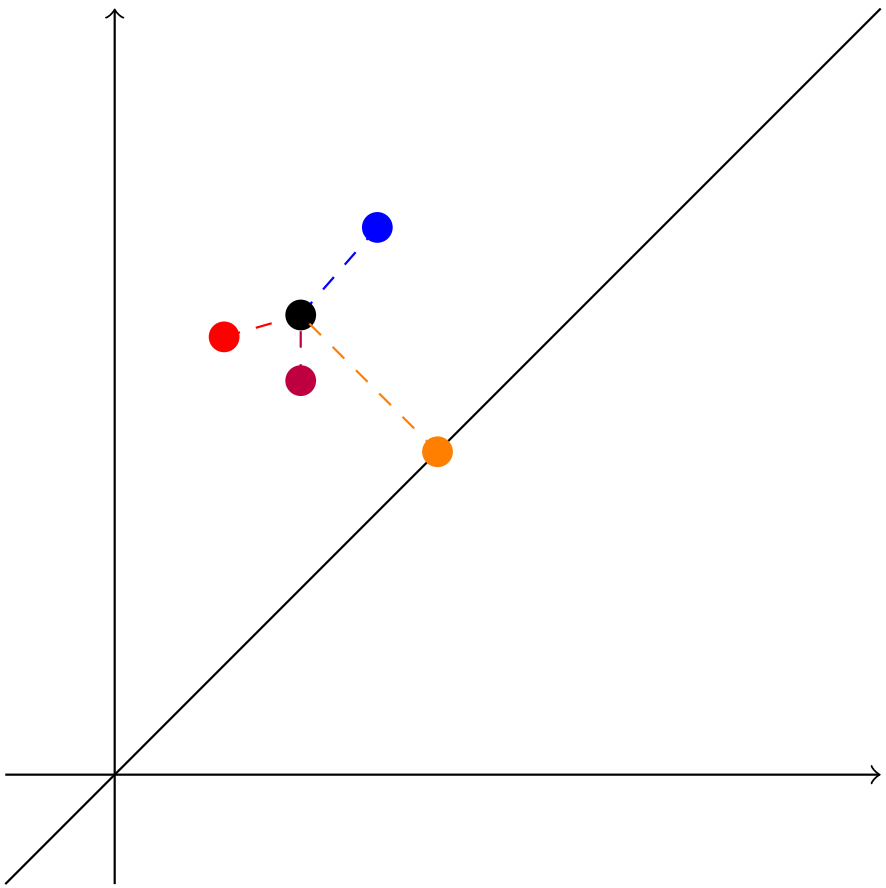
We call this (x, y) the "mean" of $(a_1, b_1), (a_2, b_2), \ldots, (a_k, b_k)$ and $n - k$ copies of the diagonal.

"Mean" of 3 points in $\mathbb{R}^2$ and 2 copies of the diagonal



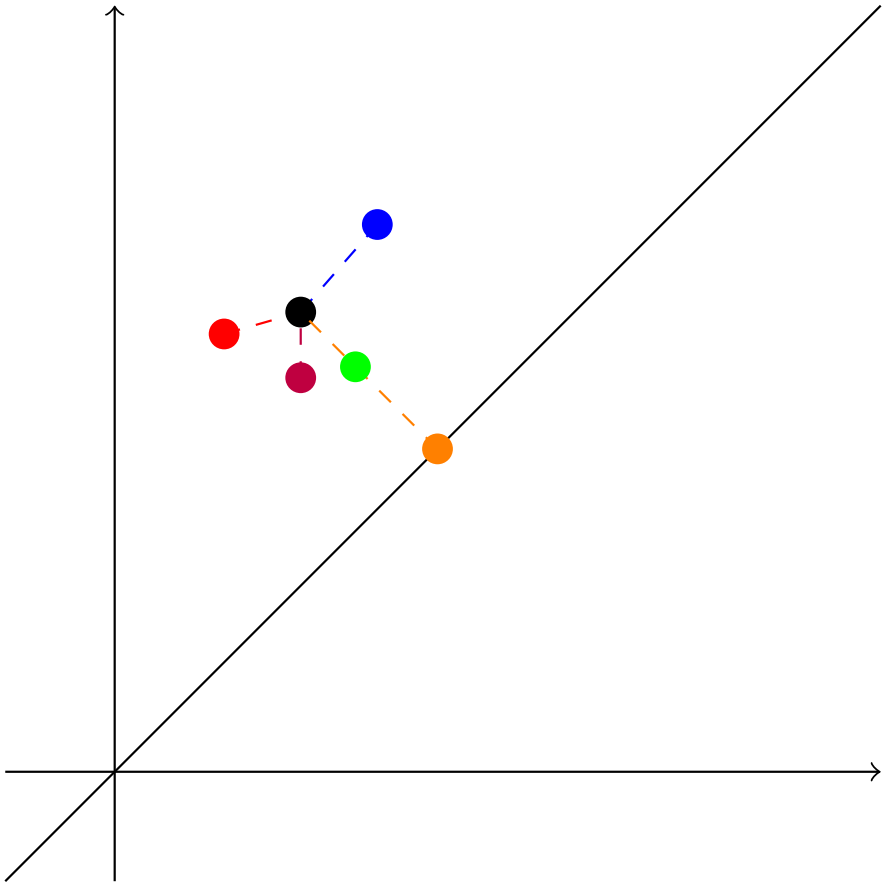
The red, blue and purple points are the (a_i, b_i). The black point is their arithmetic mean - $(\hat{x}, \hat{y})$.

"Mean" of 3 points in $\mathbb{R}^2$ and 2 copies of the diagonal



The orange is the point on the diagonal closest to the black.

"Mean" of 3 points in $\mathbb{R}^2$ and 2 copies of the diagonal



The green point is the mean of the red, blue, purple and 2 copies of the orange. It is the weighted average of the black and orange.

"Median" of points in $\mathbb{R}^2$ and copies of the diagonal

Lemma

Suppose $k > n/2$. Let $(a_1, b_1), (a_2, b_2), \ldots, (a_k, b_k)$ be points in the plane. Let (x, y) be the point in $\mathbb{R}^2$ where x is the median of $a_1, a_2 \ldots a_k$ with $n - k$ copies of $-\infty$ and y is the median of $b_1, b_2, \ldots, b_k$ with $n - k$ copies of ∞. Then (x, y) is the point in $\mathbb{R}^2$ which minimizes

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

"Median" of points in $\mathbb{R}^2$ and copies of the diagonal

Lemma

Suppose $k > n/2$. Let $(a_1, b_1), (a_2, b_2), \ldots, (a_k, b_k)$ be points in the plane. Let (x, y) be the point in $\mathbb{R}^2$ where x is the median of $a_1, a_2 \ldots a_k$ with $n - k$ copies of $-\infty$ and y is the median of $b_1, b_2, \ldots, b_k$ with $n - k$ copies of ∞. Then (x, y) is the point in $\mathbb{R}^2$ which minimizes

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

We call this (x, y) the "median" of $(a_1, b_1), (a_2, b_2), \ldots, (a_k, b_k)$ and $n - k$ copies of the diagonal if $\sum_{i=1}^{k} \|(x, y) - (a_i, b_i)\|_1 + \sum_{i=k+1}^{n} \|(x, y) - \Delta\|_1 < \sum_{i=1}^{k} \|\Delta - (a_i, b_i)\|_1$. Otherwise we say the "median" is the diagonal.

"Median" of points in $\mathbb{R}^2$ and copies of the diagonal

Lemma

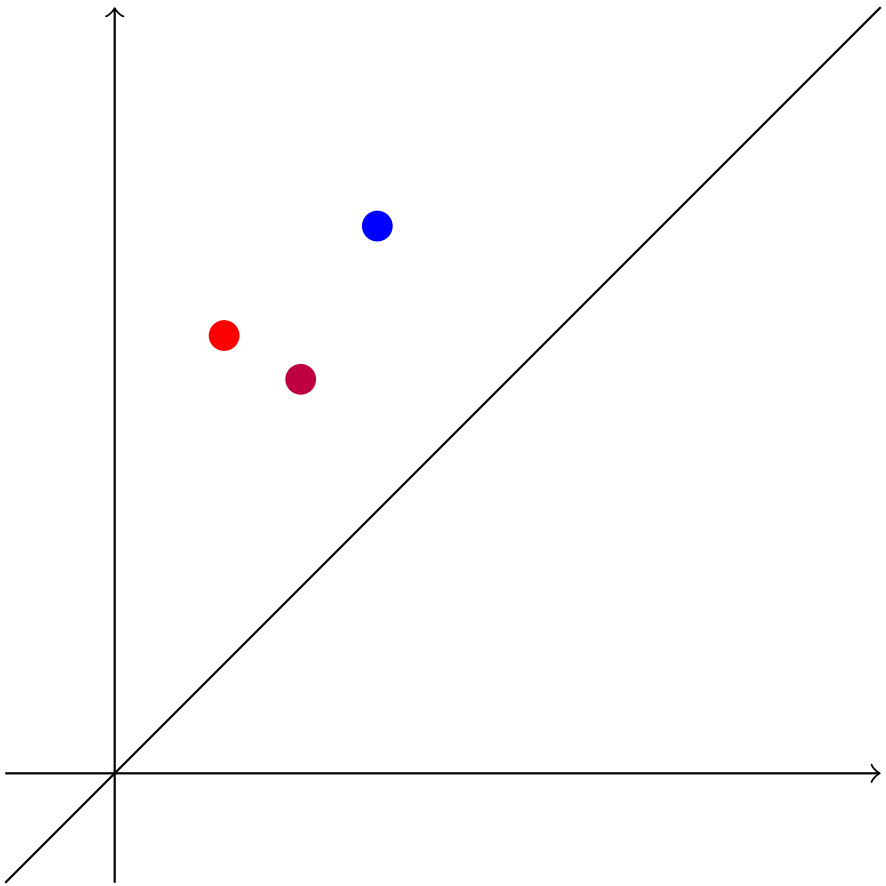
If $k < n/2$ *then*

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

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

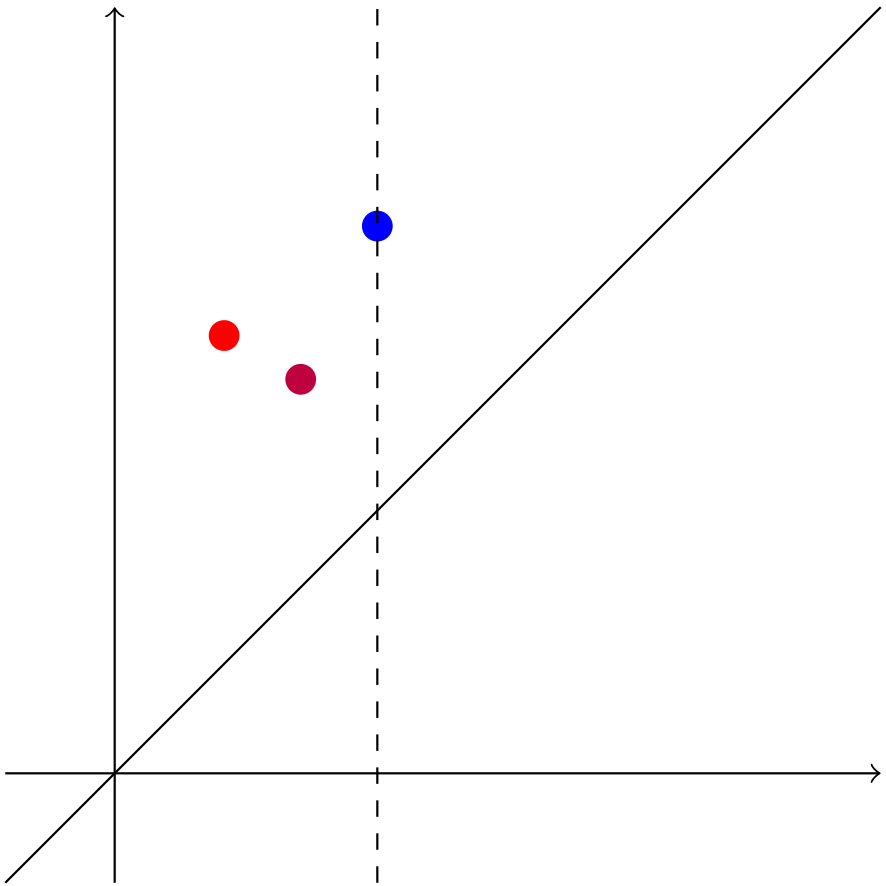
When $k < n/2$ we say that the "median" of $(x_1, y_1), (x_2, y_2), \ldots, (x_k, y_k)$ and $n - k$ copies of the diagonal is the diagonal.

"Median" of 3 points in $\mathbb{R}^2$ and 2 copies of the diagonal



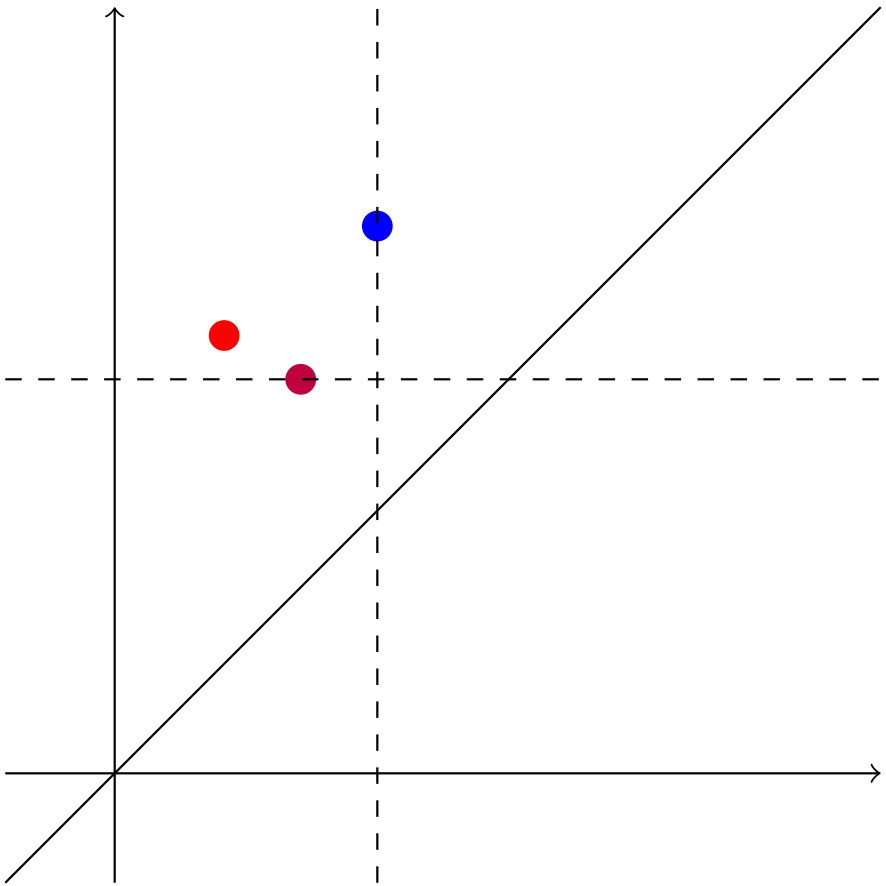
The red, blue and purple points are the (a_i, b_i).

"Median" of 3 points in $\mathbb{R}^2$ and 2 copies of the diagonal



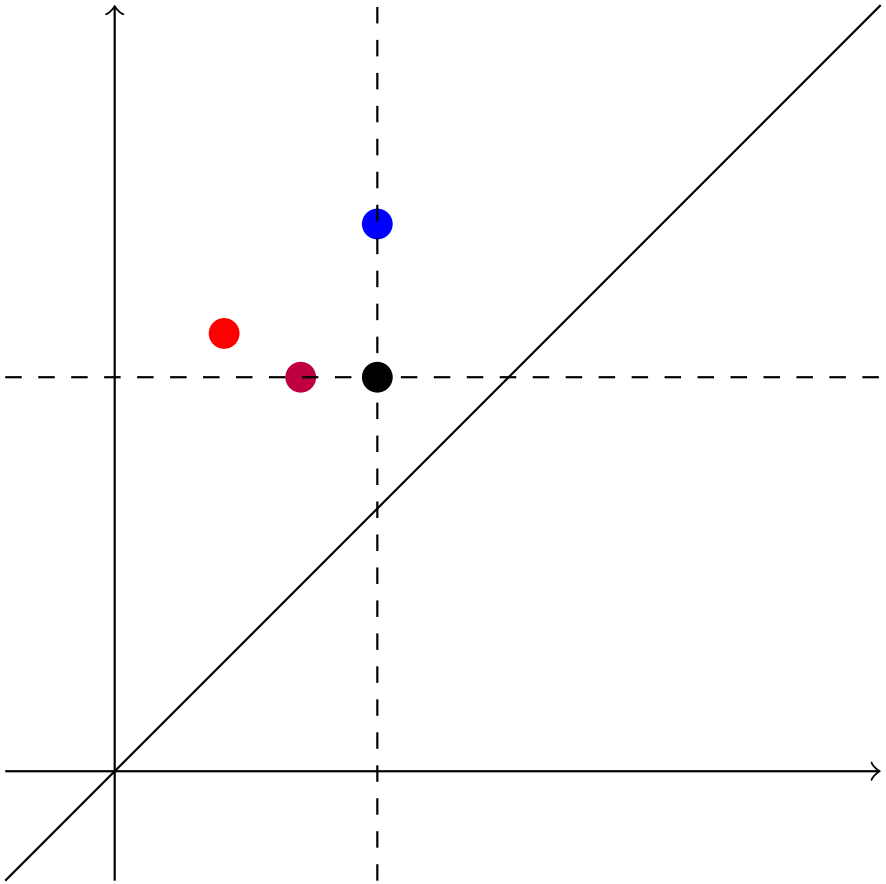
To find the x coordinate we take the median of $\{a_1, a_2, a_3, \infty, \infty\}$.

"Median" of 3 points in $\mathbb{R}^2$ and 2 copies of the diagonal



To find the y coordinate we take the median of $\{b_1, b_2, b_3, -\infty, -\infty\}$.

"Median" of 3 points in $\mathbb{R}^2$ and 2 copies of the diagonal



Selections and Matchings

Given a set of diagrams $X_1, \cdots, X_N$, a *selection* is a choice of one point from each diagram, where that point could be Δ.

Selections and Matchings

Given a set of diagrams $X_1, \cdots, X_N$, a *selection* is a choice of one point from each diagram, where that point could be Δ.

A *matching* is a set of selections so that every off-diagonal point of every diagram is part of exactly one selection.

Selections and Matchings

Given a set of diagrams $X_1, \cdots, X_N$, a *selection* is a choice of one point from each diagram, where that point could be Δ.

A *matching* is a set of selections so that every off-diagonal point of every diagram is part of exactly one selection.

Each matching Φ gives a candidate μ_Φ for the mean by taking the diagram whose points are the means of each of the selections. The mean is one of these μ_Φ so we only need to compare the $F_2(\mu_\Phi)$ over all matchings Φ.

Selections and Matchings

Given a set of diagrams $X_1, \cdots, X_N$, a *selection* is a choice of one point from each diagram, where that point could be Δ.

A *matching* is a set of selections so that every off-diagonal point of every diagram is part of exactly one selection.

Each matching Φ gives a candidate μ_Φ for the mean by taking the diagram whose points are the means of each of the selections. The mean is one of these μ_Φ so we only need to compare the $F_2(\mu_\Phi)$ over all matchings Φ.

Each matching Φ gives a candidate m_Φ for the median by taking the diagram whose points are the medians of each of the selections. The median is one of these m_Φ so we only need to compare the $F_1(m_\Phi)$ over all matchings Φ.

Description of local minimums of F_2

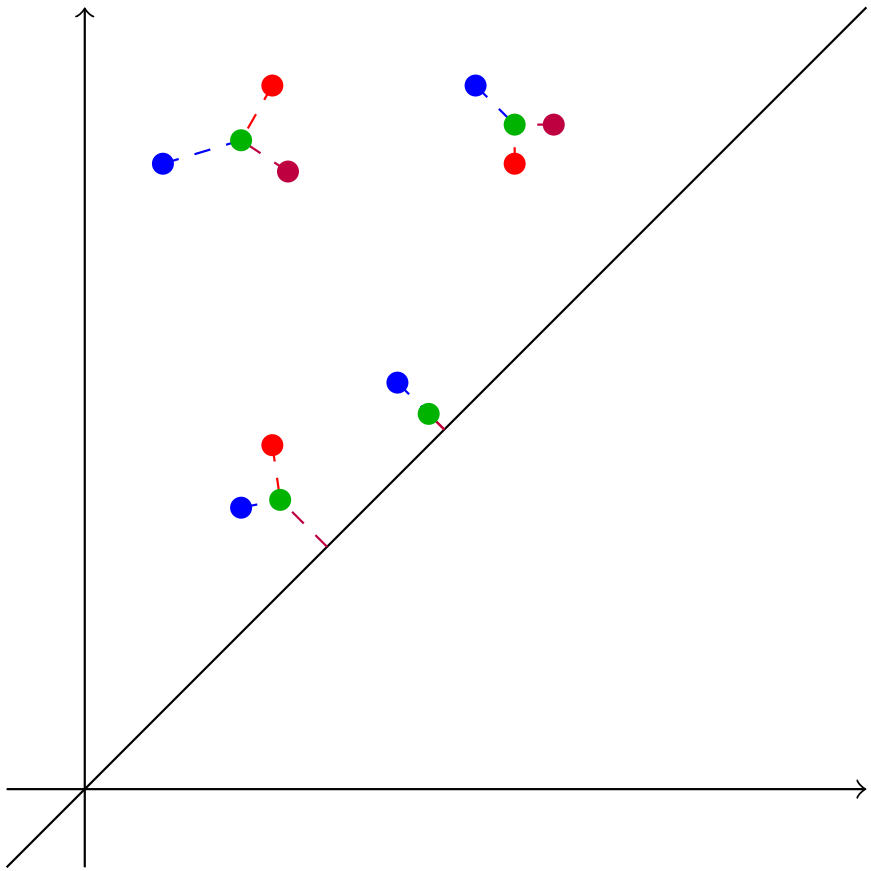
Theorem

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 ϕ_i, and each w_j is the arithmetic mean of the points $\{\phi_i(w_j)\}_{i=1,2,\ldots,m}$.

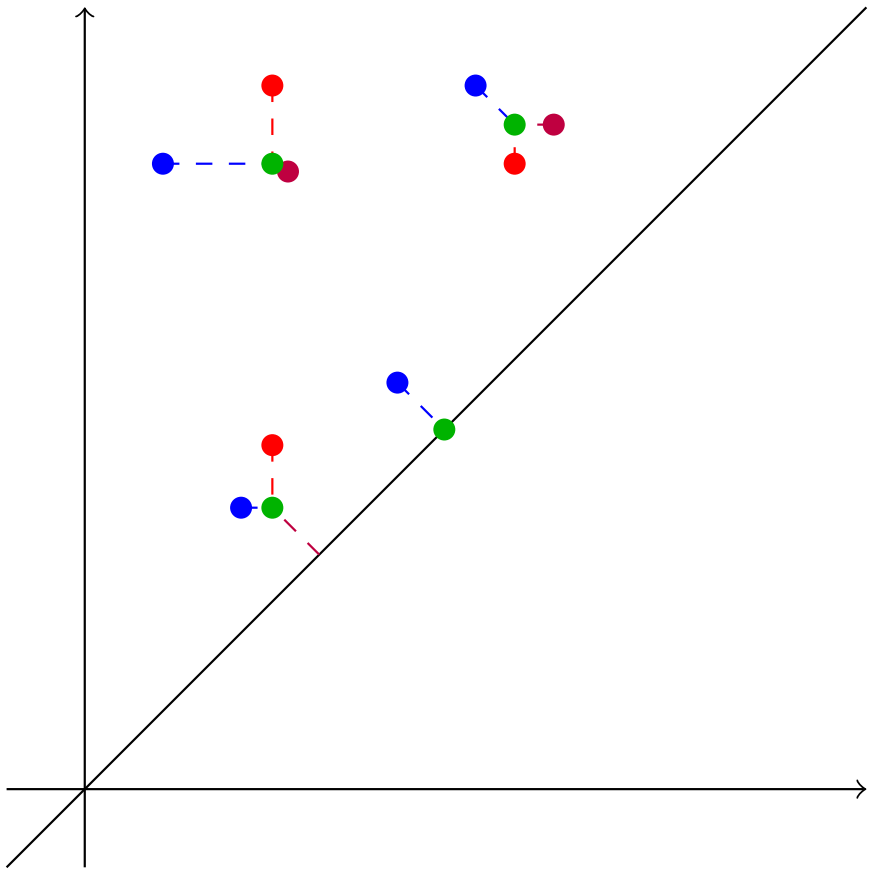
Description of local minimums of F_1

Theorem

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 and only if there whenever $\phi_i : W \to X_i$ for every optimal pairings each w_j is the "median" of the points $\{\phi_i(w_j)\}_{i=1,2,\ldots,m}$.



Mean (green) of three diagrams



Median (green) of three diagrams

The End