TDA + Statistics III: Landscape Means

In the previous two sessions, we highlighted a gap in the theory of persistent homology—the lack of unique means—and a proposed solution, persistence landscapes.

The last Jamboard finished with topological examples from Bubenik.

Here we’ll develop more of the math.
First, let's review the process of constructing them.

Given a PD: \( P = \{ (h_i, d_i): d_i \geq h_i \geq 0 \} \)

transform the persistence points using

\[
m = \frac{1}{2} (b+d), \quad h = \frac{d-b}{z}
\]

This maps the wedge \( \{ (b, d): d \geq b \geq 0 \} \)
in the first quadrant in the \( (b, d) \) plane
to the other wedge in the \( (m, h) \) plane.

Then \( (b_i, d_i) \rightarrow (m_i, h_i) = (\frac{1}{2} (b_i + d_i), \frac{1}{2} (d_i - b_i)) \)
This maps the positive $d$-axis to the diagonal $(0,d) \to (\frac{1}{2}d, \frac{1}{2}d)$ and the diagonal $d=b$ to the $m$ axis: $(b,d) \to (b,0)$.

Then we constructed peak functions $f_i$ for each $(m_i, h_i)$:
From the \( f_i \) we defined landscape functions \( \lambda_i \):

\[
\lambda_1(m) = \max \{ f_1(m), \ldots, f_k(m) \}
\]

\[
\lambda_2(m) = \text{2nd largest value of } \{ f_1(m), \ldots, f_k(m) \}
\]

\[
\lambda_j(m) = \text{\( j \)-th largest value of } \{ f_1(m), \ldots, f_k(m) \}
\]

\[
\lambda_j(m) = 0 \text{ for } j > k.
\]
This produces a decreasing sequence of functions:

\[ \lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \]

\[ \lambda_i = 0 \text{ for } i > \pi \]

We write

\[ \Delta_{\pi} = \{ \lambda_{i_1}, \lambda_{i_2}, \ldots \} \]

However, since we'll have multiple landscapes we'll write

\[ \Delta_{\pi}^j = \{ \lambda^1_j, \lambda^2_j, \lambda^3_j, \ldots \} \]
To proceed we need to introduce the concept of a norm in a vector space. This generalizes the length of a vector in $\mathbb{R}^n$.

In $\mathbb{R}^n$, $\mathbf{x} = (x_1, \ldots, x_n)$, $\|\mathbf{x}\| = \sqrt{x_1^2 + \cdots + x_n^2}$.

From multivariable calculus and linear algebra, we know:

1. $\|\mathbf{x}\| \geq 0$ and $\|\mathbf{x}\| = 0$ iff $\mathbf{x} = \mathbf{0}$, positive definite.

2. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$, triangle inequality.

3. $\|c\mathbf{x}\| = |c|\|\mathbf{x}\|$, $c \in \mathbb{R}$, absolutely homogeneous.
Def: Let $V$ be a vector space over $\mathbb{R}$. A function $\|\cdot\| : V \to \mathbb{R}$ is called a norm if it satisfies properties 1.-3.

Key Example: Let $C_{[a,b]}$ denote the vector space of continuous functions on $\mathbb{R}$. For $f \in C_{[a,b]}$, define

$$\|f\| = \int_a^b |f(x)| \, dx$$

Prop: $\|\cdot\|$ is a norm on $C_{[a,b]}$

Pf: Homework #1.
Example: Let $f$ be the peak function corresponding to a transformed persistence point $(m_0, h_0)$.

$$\|f\| = \int_a^b |f(m)| \, dm$$

$$= \int_{m_0-h_0}^{m_0+h_0} |f(m)| \, dm$$

= Area of triangle

$$= \frac{1}{2} \cdot 2h_0 \cdot h_0$$

$$= h_0^2 = \frac{1}{4} (d_0 - b_0)^2$$

$$m_0 = \frac{1}{2} (b_0 + d_0)$$

$$h_0 = \frac{1}{2} (d_0 - b_0)$$
We are, however, interested in landscapes, not peak functions.

Suppose \( \Lambda = \{ \lambda_1, \lambda_2, \ldots, \lambda_k, 0, \ldots \} \) is a persistent landscape. Each \( \lambda_i \) is continuous.

Define

\[
\| \Lambda \| = \sum_{j=1}^{\infty} \| \lambda_j \|
\]

where \( \| \lambda_j \| = \int_0^b |\lambda_j(m)| \, dm \)

\( b \) is chosen to be larger than sup of \( m \) where \( \lambda_i(m) > 0 \).
Since the landscape functions $\lambda_i \geq 0$, $||\lambda_i||$ is the area under the graph of $\lambda_i$.

Ex. Suppose $\Lambda$ is the landscape function for two persistence points $(b_1, d_1), (b_2, d_2)$, $\Lambda$ will be constructed from two triangles. A simple geometric argument shows that regardless of the values of the $b_i, d_i$:

$$||\Lambda|| = \frac{1}{4} (d_1 - b_1)^2 + \frac{1}{4} (d_2 - b_2)^2$$  (Exercise 2)
In fact this example generalizes.

Prop If \( \Lambda \) is constructed from the peak functions for \( \{(b_i, d_i), \ldots, (b_k, d_k)\} \), then

\[
\| \Lambda \| = \sum_{i=1}^{k} \frac{1}{d_i} (d_i - b_i)^2
\]

Pf (Outline) If we are given a landscape constructed from peak functions, for each point \( (m, h) \) in the first quadrant, we can assign a value \( = \# \text{ peak functions } f_i, \ h = f_i(m) \).
The diagram illustrates a function $F(m, h)$ mapping from $N \subseteq \mathbb{R}$ to another set. The function appears to be defined recursively with steps labeled 1, 2, and 3. Each step involves a transformation that is not explicitly detailed in the image. The function's behavior is suggested to be analyzed with parameters $\lambda_1$ and $\lambda_2$. The diagram suggests a complex relationship between $m$, $h$, and the resulting values of the function.
The total area under the graphs of the $\lambda_i$ (bottom) so the sum of the areas under the $\lambda_i$ can be computed from the top figure.

- Areas labeled 1 count once, since they lie "below" a single peak.
- Areas labeled 2 count twice, since they lie "below" two peaks,
- and so on.
In general, a region labeled $j$ in the top figure is contained in $j$ triangles defined by peak functions.

So this area lies under the graph of $\Lambda_j$ but not $\Lambda_{j+1}$ and will be counted in the integrals for $\|\lambda_j\|$, ..., $\|\lambda_j\|$. 
This result directly relates persistence diagrams to persistence landscape norms.

Let $V$ be a vector space spanned by the set of persistence landscapes. So every element is a sum of landscapes.

Notice, the sum of persistent landscapes is not necessarily a landscape of a persistence diagram.
Let $\mathbf{A} \in \mathbb{L}$, then

$$L = \{\lambda_1, \lambda_2, \ldots, \lambda_k, 0, \ldots\}$$

This sequence is eventually 0 because $L$ is a finite sum of landscapes of persistence diagrams all of which are 0 after a finite number of terms.

Prop

$$|\mathbf{A}| = \sum_{i=1}^{k} ||\lambda_i||$$ defines a norm on $L$.

Ref. Exercise
Now we can return to means of landscapes.

Suppose $P_j$, $j=1, \ldots, n$ are persistence diagrams. Each corresponds to a barcode, which we'll call $B_j$. A point $(b_i, d_i) \in P_j$ iff there is a bar extending from $b_i$ to $d_i$ in $B_j$.

This will be useful in interpreting means.
Let $\Lambda_{\bar{p}_j}$ be the persistence landscape of $P_j$.

Define

\[ \overline{\Lambda}_{\bar{p}_j} = \bar{\xi}_1, \ldots, \bar{\xi}_k, 0, \ldots, 0^3 \]

where

\[ \bar{\lambda}_i = \frac{1}{n} \sum_{i=1}^{n} \lambda_i^i \]

the mean of the intrinsic landscape functions of each of the $\Lambda_{\bar{p}_j}$.

The question is how to interpret the mean.
Bubenik provides the following interpretation: 

"If $B_1, \ldots, B_n$ are the bar codes corresponding to the persistence landscapes $\Lambda', \ldots, \Lambda'$, then for each $i$, 

$\bar{\lambda}_i(m)$

is the average value of the largest radius interval centered at $m$ that is contained in $i$ intervals in the barcodes $B_1, \ldots, B_n$.\"
The exercises will contain examples that explore this statement.