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 : $P = \{(b_i, d_i) : d_i \geq b_i \geq 0\}$

transform the persistence points using

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

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.

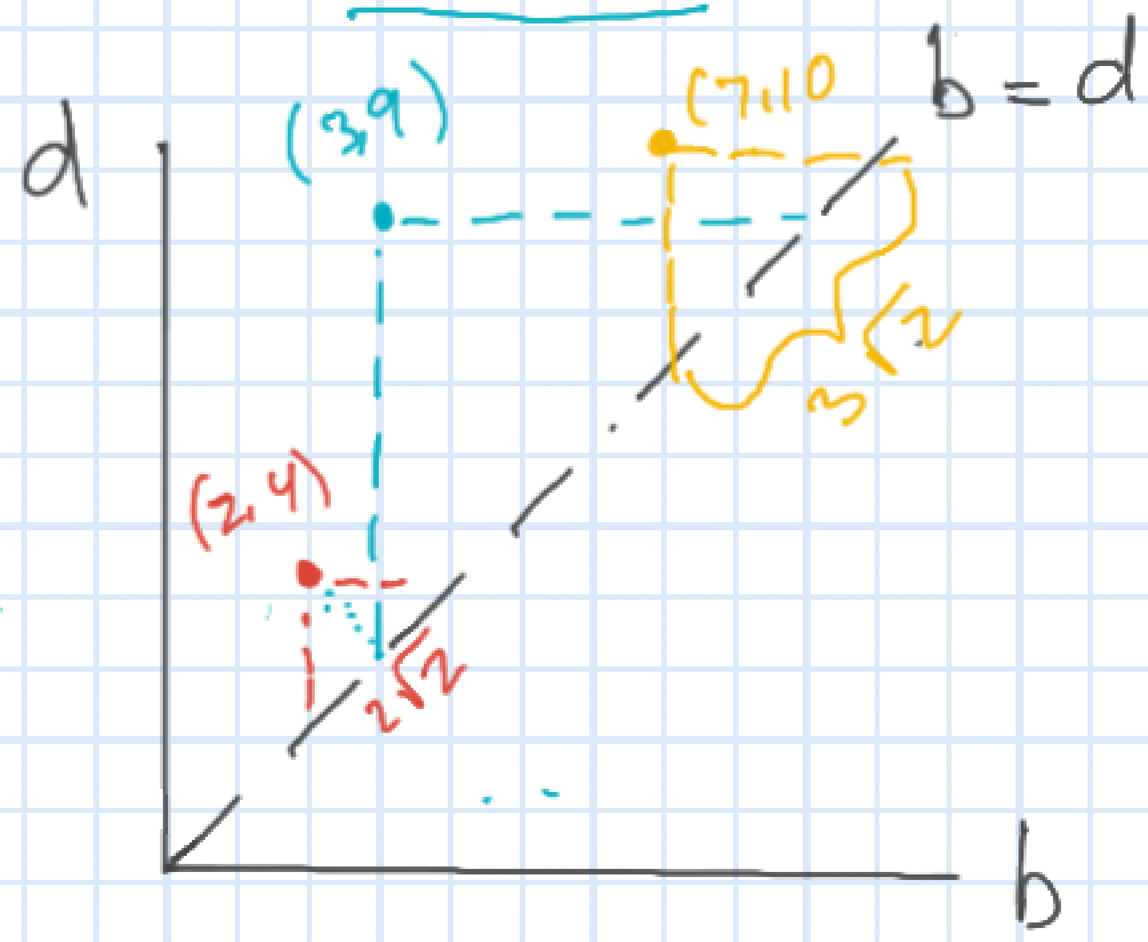
Then $(b_i, d_i) \longrightarrow (m_i, h_i) = \left(\frac{1}{2}(b_i+d_i), \frac{1}{2}(d_i - b_i)\right)$

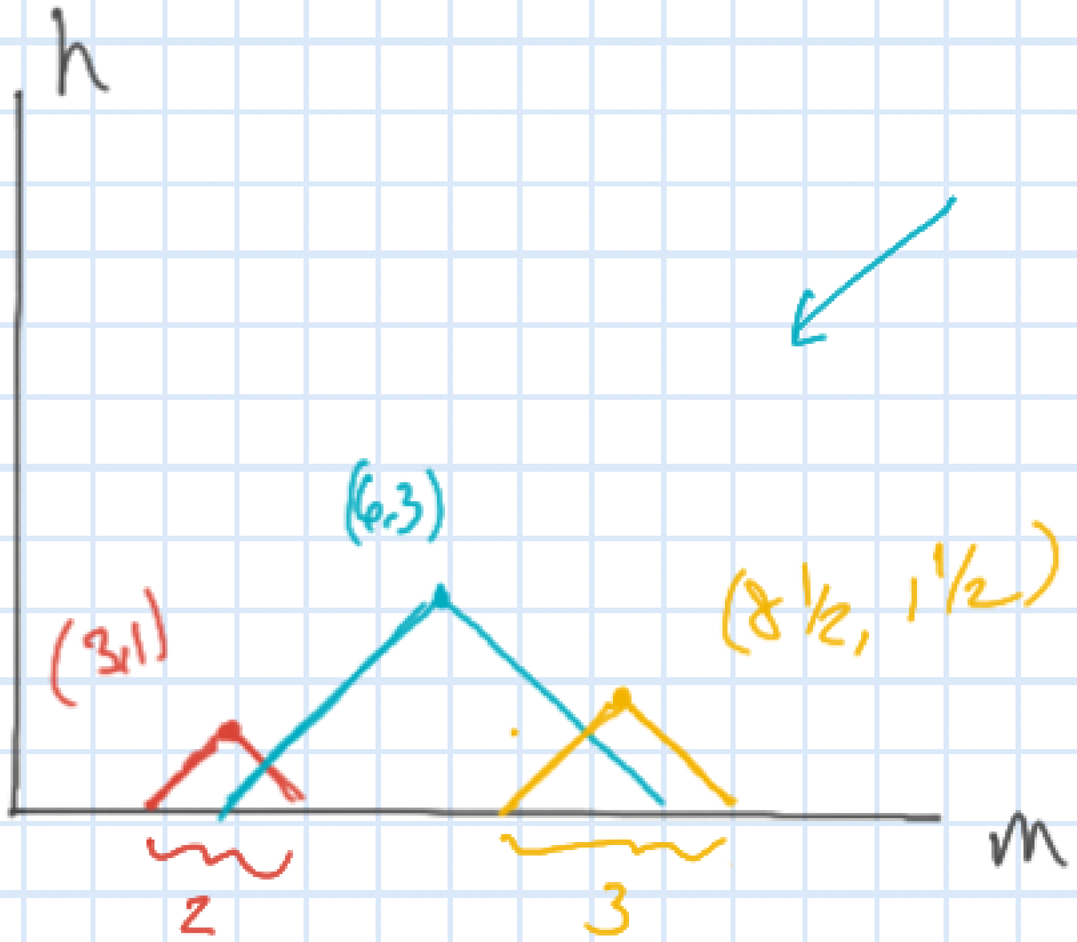
This maps the positive d-axis to the diagonal:

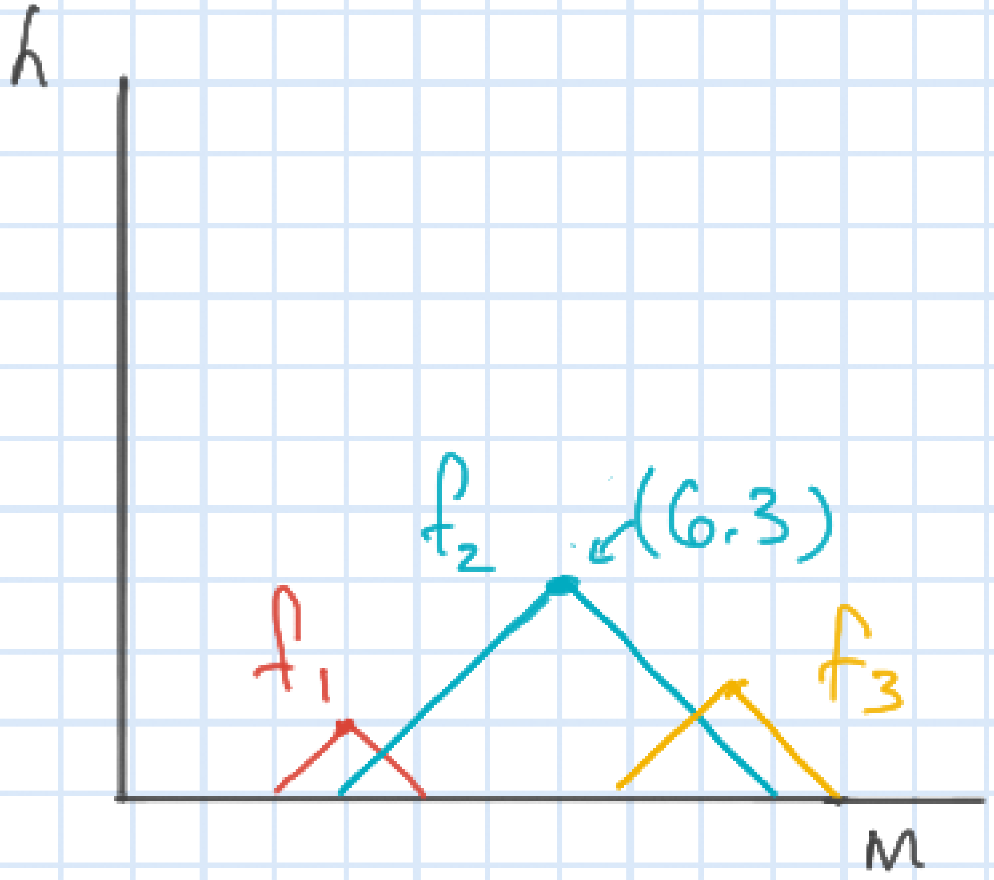
$(0,d) \to (\frac{1}{2}d, \frac{1}{2}d)$ and the diagonal $d=b$

to the m axis: $(b,d) \to (b,0)$.

Then we constructed peak functions f_i for each (m_i, h_i):







$$f_i(m) = \begin{cases} 0 & 0 \leq m \leq b_i \\ m-b & b_i \leq m \leq \frac{1}{2}(b_i+d_i) \\ d-m & \frac{1}{2}(b_i+d_i) \leq m \leq d_i \\ 0 & d_i \leq m \end{cases}$$

From the f_i we defined landscape functions λ_i

$$\lambda_1(m) = \max\{f_1(m), \dots, f_k(m)\}$$

$$\lambda_2(m) = 2^{nd} \text{ largest value of } \{f_1(m), \dots, f_k(m)\}$$

$$\vdots$$

$$\lambda_j(m) = j^{th} \text{ largest value of } \{f_1(m), \dots, f_k(m)\}$$

$$\lambda_j(m) = 0 \quad \text{for } j > k.$$

This produces a decreasing sequence of functions:

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \dots$$

$\lambda_j \equiv 0$ for $j > \# P$

We write $\Lambda_P = \{\lambda_1, \lambda_2, \dots\}$

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

$$\Lambda_{P_j} = \{\lambda_1^j, \lambda_2^j, \lambda_3^j, \dots\}$$

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 $\mathbb{R}^n$, $\bar{x} = (x_1, \dots, x_n)$, $\|\bar{x}\| = \sqrt{x_1^2 + \dots + x_n^2}$

From multivariable calculus and linear algebra we know

1. $\|\bar{x}\| \geq 0$ and $\|\bar{x}\| = 0$ iff $\bar{x} = \bar{0}$, positive definite
2. $\|\bar{x} + \bar{y}\| \leq \|\bar{x}\| + \|\bar{y}\|$, triangle inequality
3. $\|c\bar{x}\| = |c| \, \|\bar{x}\|$, $c \in \mathbb{R}$, absolutely homogeneous

Def: Let V be a vector space over $\mathbb{R}$. A function

$$\| \ \| : V \to \mathbb{R}$$

is called a norm if it satisfies properties 1.–3.

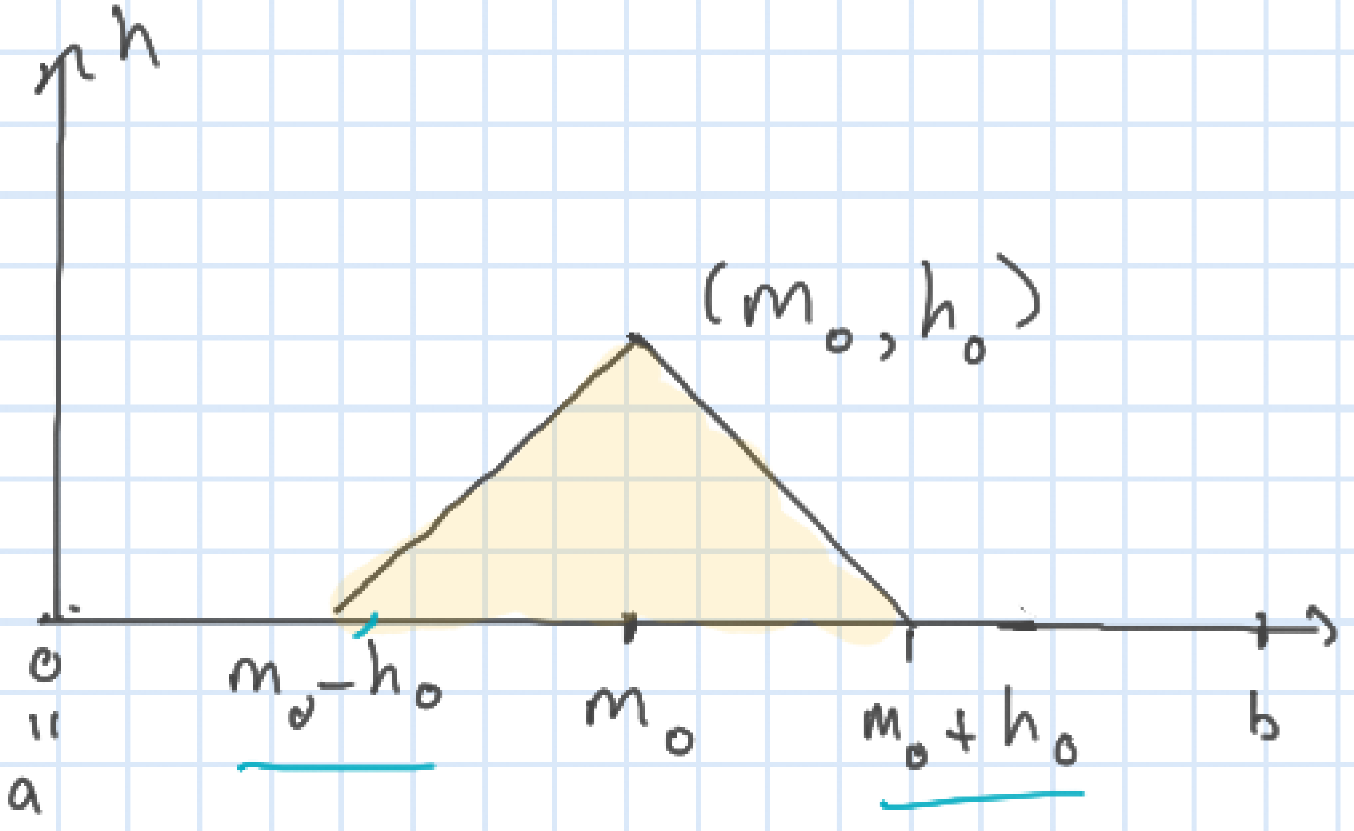
Key Example: Let $C_{[a,b]}$ denote the vector space of continuous functions on $\mathbb{R}$. For $f \in C_{[a,b]}$ define

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

Prop: $\| \ \|$ 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)



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

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

$$= \text{Area of triangle}$$

$$= \frac{1}{2} \underbrace{2h_0}_{\text{base}} \cdot \underbrace{h_0}_{\text{height}}$$

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

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

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

We are, however, interested in landscapes, not peak functions.

Suppose $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_k, 0, \ldots\}$

is a persistence landscape. Each λ_i is continuous.

Define

$$\|\Lambda\| = \sum_{j=1}^{\infty} \|\lambda_i\|$$

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

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 λ_i.

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

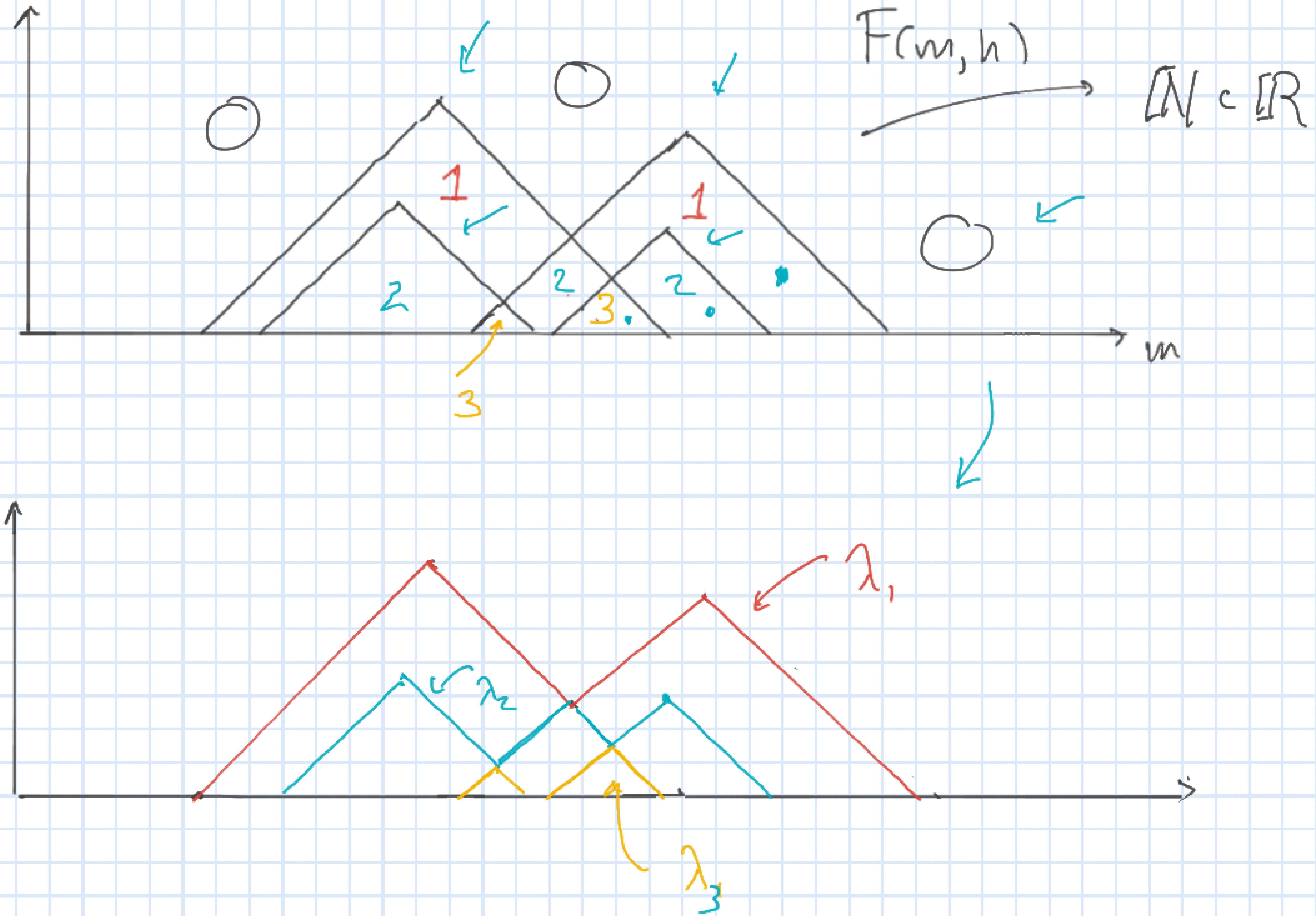
$$\|\Lambda\| = \frac{1}{4}(d_1 - b_1)^2 + \frac{1}{4}(d_2 - b_2)^2 \quad \text{(Exercise 2)}$$

In fact this example generalizes.

Prop If Λ is constructed from the peak functions for $\{(b_1, d_1), \ldots, (b_k, d_k)\}$ then

$$\|\Lambda\| = \sum_{i=1}^{k} \tfrac{1}{4}(d_i - b_i)^2$$

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 = # peak functions f_i, $h \leq f_i(m)$.


F(m, h)
ℕ ⊂ ℝ
0
0
0
1
1
2
2
2
3
3
m
λ1
λ2
λ3

The total area under the graphs of the λ_i (bottom)

— so the sum of the areas under the λ_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 λ_j but not λ_{j+1} and will be counted in the integrals for

$$\|\lambda_1\|, \dots, \|\lambda_j\|$$

✓

This result directly relates persistence diagrams persistence landscape norms.

Let L be 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 $\Lambda \in \mathcal{L}$, then

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

This sequence is eventually 0 because $\underline{L}$ is a <u>finite</u> sum of landscapes of persistence diagrams all <u>of</u> which are 0 after a <u>finite</u> # of terms.

<u>Prop</u> $\|\Lambda\| = \sum_{i=1}^{k} \|\lambda_i\|$ defines a norm on $\mathcal{L}$.

Pf. Exercise

Now we can return to means of landscapes.

Suppose P_j, $j=1,\dots n$ 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 Λ_{P_j} be the persistence landscape of P_j

Define

$$\overline{\Lambda}_{P_j} = \{\overline{\lambda}_1, \dots \overline{\lambda}_k, 0, \dots 0\}$$

where

$$\overline{\lambda}_i = \frac{1}{n} \sum_{j=1}^{n} \lambda_i^j$$

the mean of the i^{th} landscape functions of each of the $\overline{\Lambda}_{P_j}$

The question is how to interpret the mean.

Bubenik provides the following interpretation:

"If $B_1, \dots, B_n$ are the bar codes corresponding to the persistence landscapes $\Lambda^1, \dots, \Lambda^n$, 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 $B_1, \dots, B_n$

The exercises will contain examples that explare this statement.