Geometric and Topological Shape Analysis

Investigating and summarising the shape of data

Yossi Bokor

A thesis submitted in fulfillment of
the requirements for the degree of
Doctor of Philosophy

Mathematics
Australian National University & The University of Sydney

2023

Abstract

Tools from geometry and topology can be applied in a wide variety of settings. In particular, they are adept at exploring and summarising the shape of data. By considering the shape of objects, we can answer questions related to object reconstruction and classification problems. These problems often face difficulty distinguishing signals from noise, which can be overcome using ideas from persistent homology and computational geometry. We first look at an application of geometry and topology to learning the abstract structure of embedded stratified spaces and then consider an object classification problem relating to human mesenchymal stem cells.

Statement of originality and authorship attribution

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Chapter 1 of this thesis is published as [5]. I designed the algorithms for learning the abstract structure, performed the numerical simulations, and wrote the drafts of Sections 1.1, 1.2, 1.3.

Chapter 3 analyses human mesenchymal stem cells, purchased from Lonza (catalogue #: PT-2501), cultured, fixed, stained, and imaged by Dr Florian Rehfeldt (University of Bayreuth). Information was extracted from the images by both of us. The analysis used code I developed with Dr Patrice Koehl (University of California, Davis).

Yossi Bokor, 28th October, 2022.

Acknowledgements

I acknowledge the traditional owners of the lands on which I have written this thesis, the Gadigal people of the Eora nation and the Ngunnawal people, and the Elders, past and present. I acknowledge the ongoing connection that Aboriginal people have to these lands and recognise Aboriginal people as the original custodians of these lands.

The list of people I want to thank is (thankfully) countable, but in the moment of writing this, feels infinite, so the following is a non-exhaustive collection of acknowledgements.

My supervisors, Kate, Vanessa, Vigleik, Lamiae, Jonathan & Stephan have given me so much of their time. Without them, this thesis would not exist, and I would know a small fraction of the things I do.

My parents, Imi and Bea, have always supported me in my studies, and again, without their support, I would not have written this thesis. They have always encouraged me to follow my interests, and despite what everyone assumes, never pressured me to study mathematics. My aunt, Vera, and grandmothers, Marianne and Agnes, have similarly encouraged and supported me.

My office mates over the years, Ivo, Jaklyn, James, Grace, Joe, Oliver, Abi, Mark & Kyle.

My collaborators Chris, Florian, Patrice, Stephan, Fariza, Benjamin, Martin, Vidit & Primoz have helped me in many ways.

Now, the hard part, the list of friends, which I will only start: Adam, Clancy, Hannah, Willem & James, Feng, Ellen, Liliana, Sophia, Lucy, Cynthia, Miranda, Kusum, Sanjana, Annie, Wayne, Wiwid, Rodd, Jen, Peter, Jordan, Cate, Jack, Ash & Alice, Max, Maca & Matthew & Maeve, Cameron, Sarah, Adeline, William, Ross, Steven, Nick, Camille, Tyson & Charles, David & Mel & Jacob, Harry, Paul, Jieyang, Lucas, Pat, . . .

Much of where and who I am today, my attitudes, my behaviours, and my ability to complete a thesis, are consequences of interactions with many people. Some of these people have been named above. But so many have not, and I would never be able to name them all, as I either never knew their name, or I have forgotten it. I have had many jobs across a variety of industries over the years, and I regularly am reminded about how much I grew with each job. In particular, the kids I taught and coached swimming, as well as the people I worked with during my time at ADAHC, have imparted an understanding of the importance of enjoying the small things and moments in life. These moments can be fleeting, but they can also be a thread throughout a period of time. Without these moments, I would not have finished a thesis. And so, this thesis is dedicated to the small moments in life, and the people (known and unknown) you share them with.

I would also like to thank the examiners for their helpful comments.

Contents

Introduction

> But in the end it's only a passing thing, this shadow; even darkness must pass.
>
> J. R. R. Tolkien, *The Lord of the Rings*

This thesis can be split into two parts, with Chapters 1 and 2 forming the first, and Chapter 3 the second. Chapter 1 presents an algorithm for learning the abstract structure underlying an embedded graph $|G| \subset \mathbb{R}^n$ given an ε-sample P of $|G|$, as well as a method for modelling the embedding. Learning the abstract structure is extended to embedded 2-complexes $|X| \subset \mathbb{R}^n$ from ε-samples P in Chapter 2. Chapter 3 is a change in topic and pace, which explores the use of persistent homology and persistence diagrams to identify abnormal growth patterns in cultured human mesenchymal stem cells.

While these topics seem disconnected, they can be considered two sides of the same dice. In both settings, we are using tools from geometry and topology to understand the *shape* of the data at hand. In Chapters 1 and 2 the end goal is to understand the abstract structure underlying a point cloud, while in Chapter 3 we use shape to identify subpopulations.

All three chapters aim to be self-contained, and hence there is no background chapter in this thesis, and readers will find definitions from Chapter 1 repeated in Chapter 2.

People familiar with topological data analysis, computational topology and computational geometry will notice that Chapters 1 and 2 are centred on results that could be considered purely problems in computational geometry. This is in part because when working with data sets, topological data analysis often becomes computational geometry. In particular, this occurs when we care about the extrinsic structure of the data within the ambient space, rather than just the intrinsic topology and geometry. Exploring the computational geometry aspects of these problems comes with benefits as well as costs. While we are able to develop an algorithm that is more efficient in terms of time complexity, often the associated proofs about its correctness are, to quote a supervisor, 'gross'. This grossness stems from the expressions which bound distances between various objects, as they do not simplify to a 'nice, neat, expression'. These expressions could be approximated to become 'nicer', without affecting the proofs, but then the assumptions placed upon our underlying spaces become stricter.

Chapter 3 explores the use of geometric and topological tools to understand growth patterns in human mesenchymal stem cells. In particular, we use morphological features of the cells to identify subgroups within experimental conditions. By identifying the sub-populations, we are able to clean the data when investigating the impact of environmental conditions on multi-potent cells in the future, as well as potentially having an unbiased method for identifying unipotent cells. The cells were purchased, cultured and imaged by Dr Florian Rehfeldt (Bayreuth University).

CHAPTER 1

Learning and Modelling Graphs

> Instead of a stable truth, I choose unstable possibilities.
>
> Haruki Murakami, *Killing Commendatore*

Increases in the quantity and complexity of collectable data have led to the search for new methods for efficiently discovering and modelling their underlying structures. The importance of dimensionality reduction of large amounts of data grows with the embedding dimension. By expanding the class of underlying structures which can be detected and modelled, we aim to address some of the difficulties. To improve dimensionality reduction's efficiency and accuracy, we remove the manifold assumption where the dimension is constant and instead treat it as a stratified space, learning the local dimension in the algorithm. We focus on one-dimensional stratified spaces (i.e. graphs) and here provide a new method for dimensionality reduction and compression.

Manifold learning is a method of detecting and modelling structures underlying data sets. There are numerous algorithms and theorems for learning geometric and topological features of manifolds from (noisy) samples, such as dimension or the manifold itself (see [10], [13], [14]). These algorithms make assumptions about the manifold and the sampling procedure, often in the form of curvature restrictions and conditions on the sample's density and noise. Unfortunately, these assumptions are not satisfied by point clouds arising in many applications, such as geospatial transportation network data of vehicle movement. We move towards resolving this problem by expanding the set of allowable underlying structures to include stratified spaces. A *stratified space* is a space described by gluing together (manifold) pieces, called strata. There are no restrictions placed upon each stratum's dimension, and the gluing can give rise to a variety of interesting and complex local structures.

Bendich et al. ([2], [3]) describe an algorithm which, under certain conditions, can identify if two points have been sampled from the same stratum of a stratified space. This algorithm does not provide a method for learning the global abstract structure. In related work, Stolz et al. ([22]) present an algorithm for detecting when points have been sampled from two intersecting manifolds which is a cruder splitting than the splitting into stratified subspaces. They have some experimental verification but no theoretical guarantees.

The closest previous work to this paper is [1], in which Aanjaneya et al. consider reconstructing *metric graphs* to detect branch points and the graph structure. There are a few crucial differences. They focus in on the reconstruction of the metric, with input intrinsic distances on the metric graph (plus noise) and the aim to reconstruct a metric graph that is homeomorphic and close as metrics. This means that the theoretical guarantees are about the lengths of edges in the metric graph instead of geometric conditions on an embedding. Crucially, they do not need to consider vertices of degree 2 as in a metric space setting these are points on an edge.

In contrast, this chapter describes an algorithm for modelling a linear embedding of a simple graph from a point cloud sample and provide theoretical guarantees in terms of the geometric embedding that the graph structure modelled is equivalent to the structure embedded.

Definition 1.0.1 (Graph)**.** *A* graph *G consists of*

1. *a set of vertices $V = \{v_i\}_{i=1}^{n_v}$,*
2. *a set of edges $E = \{(v_{j_1}, v_{j_2})\}_{j=1}^{n_e}$.*

For any graph G, the boundary operator *$\partial_G : E \to V \times V$, maps an edge to the two boundary vertices. We can represent ∂_G via the* boundary matrix *B, which is the $n_v \times n_e$ matrix with $B[i, j] = 1$ if $v_i = v_{j_1}$ or $v_i = v_{j_2}$, and 0 otherwise. Edges (v_{j_1}, v_{j_2}) are* open*, and their boundary consists of the two vertices.*

Given a graph G, we can embed it into $\mathbb{R}^n$ in numerous ways. We will restrict to linear embeddings, such that at degree 2 vertices, the angle between edges is not π.

Definition 1.0.2 (Linear embedding). *A linearly embedded graph*

$$|G|= (G, \phi_G) \subset \mathbb{R}^n$$

is a graph G, and a map $\phi_G : G \to \mathbb{R}^n$, such that

1. *on the vertex set V, ϕ_G is injective, and we denote $\phi_G(v)$ by v,*
2. *on E, ϕ_G is defined by linear interpolation: the embedding of an edge (u, v) is the line segment joining $\phi_G(u)$ and $\phi_G(v)$, denoted $\overline{\phi_G(u)\phi_G(v)} = \overline{uv}$,*
3. *embedded edges $\overline{uv}, \overline{u'v'}$ only intersect if they share a boundary vertex, say $v' = v$, and their intersection is $\phi_G(v)$.*

We restrict our attention to embedded graphs $|G|$ such that at a degree two vertex v, the embedded edges, say $\overline{uv}$, $\overline{wv}$ form an angle $\alpha \neq \pi$.

Please note that with an abuse of notation we will usually use v to denote both the abstract vertex and the embedded location $\phi_G(v)$, and use $\overline{uv}$ to denote both the abstract edge and the embedded image of that edge by ϕ_G. It should always be clear from context whether we are referring to an element in the abstract structure or to its image in $\mathbb{R}^n$.

Throughout this paper, we use the following conventions. For two points $x, y \in \mathbb{R}^n$, $\|x-y\|$ is the distance between x and y in the standard Euclidean metric on $\mathbb{R}^n$, $\langle x, y\rangle$ is the inner product of x and y. For a point $x \in \mathbb{R}^n$ and a set $Y \subset \mathbb{R}^n$, we set

$$d(x, Y) := \inf_{y\in Y} \|x - y\|,$$

and for two sets $X, Y \subset \mathbb{R}^n$,

$$d(X, Y) := \inf_{x\in X, y\in Y} \|x - y\|.$$

Given a point $x \in |G|$, we can determine if x is on an edge, or is a vertex by considering the intersection of $|G|$ with a small ball around x. Consider $B_r(x)$ for small $r > 0$. If x is a vertex, r is less than $\|x - w\|$ for all vertices $w \neq x$ and there are no edges $\overline{uw}$ within r of x, then $B_r(x) \cap |G|$ is connected, and for each edge containing x, there is a unique point in $\partial B_r(x)$. If x is a degree 2 vertex, let the two points on $\partial B_r(x)$ be p and q, then $\angle pxq < \pi$. Now consider $x \in \overline{uv}$ for some embedded edge $\overline{uv}$, and take $r < \min\{\|x-v\|, \|x-u\|\}$. If there is some edge $\overline{wz}$ with $d(x, \overline{wz}) \leq r$, then $B_r(x) \cap |G|$ is disconnected. Otherwise, $B_r(x) \cap |G|$ is connected, and there are two points q, p in $\partial B_r(x) \cap |G|$, and $\angle pxw = \pi$. This is an adaption of the local homology of $|G|$ at x.

We suppose that we do not have the entire embedded graph $|G|$, but only a finite sample P. Furthermore, we expect noise so that $P \not\subseteq |G|$, and we can only make statements about the distance between P and $|G|$. We restrict to sufficiently dense samples P of $|G|$ with bounded noise. Let $d_H(X, Y)$ be the Hausdorff distance between two subsets X, Y of $\mathbb{R}^n$,

$$d_H(X,Y) := \max\left\{\sup_{x\in X} d(x,Y), \sup_{y \in Y} d(y,X)\right\}.$$

We consider ε-*samples* of embedded graphs $|G|$.

Definition 1.0.3 (ε-sample)**.** *Let* $|G| \subset \mathbb{R}^n$ *be an embedded graph. An* ε-sample P *of* $|G|$ *is a finite subset of* $\mathbb{R}^n$ *such that* $d_H(|G|, P) \leq \varepsilon$.

We can now state the aim of this chapter: given an ε-sample P of a linearly embedded graph $|G|$, we want to 1) detect the graph structure G, and then 2) model ϕ_G. This is a semi-parametric problem: the parameters we need to learn are the number of vertices, the number of edges, and the boundary operator. To do so, we need to decide if p is near a vertex v or far away from all vertices for each $p \in P$. This partitions our sample P into two subsets, which intuitively are P_0 containing samples p which are near a vertex, and P_1 containing samples p which are not near any vertex. We define P_0 and P_1 rigorously in Definition 1.2.5. In the process of partitioning P, we approximate the local homology at each $p \in P$ using radius r. This requires choosing a scale for approximating $|G|$ from P. The clusters in P_0 and P_1 correspond to vertices and edges in G respectively, and we can use the minimal distance between clusters in P_1 and P_0 to learn the boundary operator. Using this information, we model the embedding ϕ_G.

A necessary but not sufficient condition for a point p to be near a vertex is $B_r(p) \cap |G|$ being connected. If it is disconnected, p is not near any vertex, and if it is, we need to check the number of connected components in $B_r(p) \cap |G|$ to determine if p is near a vertex or not. As p is within ε of $|G|$, r must be greater than ε to ensure $B_r(p) \cap |G|$ is non-empty.

Fix $R > \varepsilon$. We first want to approximate $B_R(x) \cap |G|$, and then $\partial B_R(x) \cap |G|$ from P. We can approximate $B_R(p) \cap |G|$ by considering samples $q \in P$ with $\|p - q\| \leq R$. As P is an ε-sample of $|G|$, we can approximate $\partial B_R(p) \cap |G|$ by considering the samples in a spherical shell $S^{R+\varepsilon}_{R-\varepsilon}(p)$ of inner radius $R - \varepsilon$, outer radius $R + \varepsilon$ around p.

We model ϕ_G by aiming to reconstruct a probability measure ν which is supported on $|G| \subset \mathbb{R}^n$. As recorded data has errors, we cannot directly reconstruct ν, but instead construct an approximating measure ν_δ such that ν_δ is equivalent to the Lebesgue measure, and $\text{supp}(\lim_{\delta \to 0} \nu_\delta) = |G|$. We form ν_δ from a categorical mixture model of measures over the individual strata pieces, with latent variables for strata assignment. We use a Gaussian convolution for each individual strata piece to form our approximation of ν with ν_δ. We derive a log-likelihood function which is maximised through an Expectation-Maximisation algorithm (Algorithm 3).

In Section 1.1, we present and prove some geometric lemmas used throughout Sections 1.2 and 1.3, then in Section 1.2 we define (R, ε)-local structure, describe the (R, ε)-local structure of a vertex and of an edge, before providing conditions under which we can guarantee what (R, ε)-local structure a sample p has. Section 1.3 presents the algorithm, relates it to the (R, ε)-local structure, before proving that the abstract graph identified is equivalent to the original one. Finally, Section 1.4 describes the modelling process used, and contains some simulations.

1.1. Some Geometric Lemmas

As motivation for the formulas both in the definitions of local structure and the geometric assumptions of the graphs' embedding, we first prove some geometric lemmas. Throughout our process, we consider $\langle x_1 - p, x_2 - p \rangle$ for p, x_1, x_2 samples, and $\|p - x_1\|, \|p - x_2\| \in [R - \varepsilon, R + \varepsilon]$. In particular, if there are two clusters of points in the spherical shell around a sample p, all points (including p) are within ε of an edge $\overline{uv}$, and x_1 and x_2 are from different clusters, we wish to bound $\langle x_1 - p, x_2 - p\rangle$ from above.

Lemma 1.1.1. *Fix $R > 12\varepsilon > 0$ and consider a sample p within ε of an edge $\overline{uv}$. Let H be the hyper-plane through p perpendicular to $\overline{uv}$. Now take x_1, x_2 within ε of edge $\overline{uv}$ such that $\|p - x_1\|, \|p - x_2\| \in [R - \varepsilon, R + \varepsilon]$ and x_1, x_2 are on different sides of H. Then*

$$\langle x_1 - p, x_2 - p\rangle \leq -R^2 + 2R\varepsilon + 7\varepsilon^2.$$

Proof. By assumption $\|x_1 - p\|, \|x_2 - p\| \geq R - \epsilon$. As x_1, p, x_2 are all within ε of $\overline{uv}$ we know that $\angle(x_1 p x_2) \in [\pi - 2\arccos(\frac{2\epsilon}{R-\epsilon}), \pi]$. Together we can bound

$$\begin{aligned}\langle x_1 - p, x_2 - p\rangle &= \|x_1 - p\|\|x_2 - p\| \cos\angle(x_1 p x_2)\\ &\leq (R-\epsilon)^2 \cos\left(\pi - 2\arccos\left(\frac{2\epsilon}{R-\epsilon}\right)\right)\\ &\leq (R-\epsilon)^2\left(2\frac{(2\epsilon)^2}{(R-\epsilon)^2} - 1\right)\\ &\leq -R^2 + 2R\varepsilon + 7\varepsilon^2.\end{aligned}$$

□

We want to distinguish points very close to a vertex of degree 2 as close to a vertex, from points on an edge. This requires an upper bound on the angle at any vertex of degree 2 within our geometric assumptions due to the noise in sampling. The following geometric lemma motivates the upper bound given in the next section.

Lemma 1.1.2. *Fix $R \geq 12\varepsilon > 0$. Take $u, v, w \in \mathbb{R}^n$, and consider the line segments $\overline{uv}, \overline{wv}$.*

Let $p, x_1, x_2 \in \mathbb{R}^n$ be such that p and x_1 are within ε of $\overline{vw}$, x_2 is within ε of $\overline{uv}$, and $\|x_1 - p\|, \|x_2 - p\| \in [R - \varepsilon, R + \varepsilon]$.

If either

1. $\|p - v\| < 4\varepsilon$ *and*

$$\pi/2 < \angle uvw < \pi - \arctan\left(\frac{R+3\varepsilon}{6\varepsilon}\right) + \arcsin\left(\frac{R^2 - 4R\varepsilon - 9\varepsilon^2}{(R+\varepsilon)\sqrt{R^2+6R\varepsilon+34\varepsilon^2}}\right),$$

OR

2. $\|p - v\| < (R-\varepsilon)/2$ *and* $\angle uvw \leq \pi/2$

then

$$\langle x_1 - p, x_2 - p\rangle > -R^2 + 2R\varepsilon + 7\varepsilon^2.$$

Proof. Let $\widetilde{p}, \widetilde{x_1}, \widetilde{x_2}$ be the projections of p, x_1, x_2 to $\overline{uv} \cup \overline{wv}$. Without loss of generality, we assume $\widetilde{p}, \widetilde{x_1} \in \overline{wv} \cup v$, and $\widetilde{x_2} \in \overline{uv}$. Then there are $e_p, e_1, e_2 \in \mathbb{R}^n$ with $\|e_q\|, \|e_1\|, \|e_2\| \leq \varepsilon$ and

$$p = \widetilde{p} + e_p$$
$$x_1 = \widetilde{x_1} + e_1$$
$$x_2 = \widetilde{x_2} + e_2.$$

Now consider the vectors $x_1 - p$ and $x_2 - p$, we have:

$$\langle x_1 - p, x_2 - p\rangle = \langle \widetilde{x_1} - \widetilde{p}, \widetilde{x_2} - \widetilde{p}\rangle + \langle \widetilde{x_1} - \widetilde{p}, e_2\rangle - \langle \widetilde{x_1} - \widetilde{p}, e_p\rangle + \langle e_1 - e_p, x_2 - p\rangle. \tag{1.1}$$

We know that e_p is perpendicular to $\overline{vw}$ and thus it is also perpendicular to $\widetilde{x_1} - \widetilde{p}$, implying $\langle \widetilde{x_1} - \widetilde{p}, e_p\rangle = 0$. Further, we know that

$$\|\widetilde{x_1} - \widetilde{p}\| \leq \|x_1 - p\| \leq R + \varepsilon$$

as distances can only decrease when projecting onto $\overline{vw}$. Hence, to bound $\langle \widetilde{x_1} - \widetilde{p}, e_2\rangle$ we first split $e_2 = e_2' + e_2''$ where e_2' is the projection of e_2 into the plane spanned by $\overline{vu}$ and $\overline{vw}$. Note that e_2'' is perpendicular to $\widetilde{x_1} - \widetilde{p}$ and hence $\langle \widetilde{x_1} - \widetilde{p}, e_2\rangle = \langle \widetilde{x_1} - \widetilde{p}, e_2'\rangle$. From here, we need to split the proof into the two scenarios.

Assume we are in scenario 1: $\|p - v\| < 4\varepsilon$ and

$$\pi/2 < \angle uvw < \pi - \arctan\left(\frac{R+3\varepsilon}{6\varepsilon}\right) + \arcsin\left(\frac{R^2 - 4R\varepsilon - 9\varepsilon^2}{(R+\varepsilon)\sqrt{R^2+6R\varepsilon+34\varepsilon^2}}\right).$$

The angle between e_2' and $\widetilde{x}_1 - \widetilde{p}$ is either $\angle uvw + \pi/2$ or $\angle uvw - \pi/2$. Recall that we assumed $\angle uvw \in (\pi/2, \pi)$, so

$$\cos(\angle uvw - \pi/2) > 0 > \cos(\angle uvw + \pi/2)$$

and

$$\langle \widetilde{x}_1 - \widetilde{p}, e_2 \rangle = \langle \widetilde{x}_1 - \widetilde{p}, e_2' \rangle \geq \|\widetilde{x}_1 - \widetilde{p}\| \|e_2'\| \cos(\angle uvw + \pi/2) \geq -\varepsilon(R+\varepsilon) \sin \angle uvw. \tag{1.2}$$

Combining (1.1) and (1.2) we see

$$\langle x_1 - p, x_2 - p \rangle \geq \langle \widetilde{x}_1 - \widetilde{p}, \widetilde{x}_2 - \widetilde{p} \rangle - \sin \angle uvw (R+\varepsilon)\varepsilon - (R+\varepsilon)2\varepsilon. \tag{1.3}$$

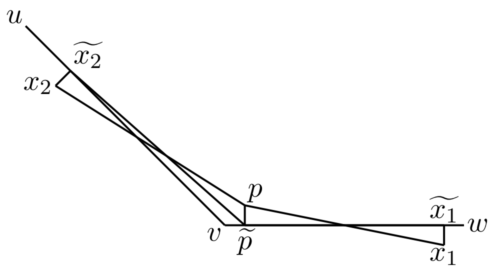


FIGURE 1.1. An example of scenario 1.

To bound $\langle \widetilde{x}_1 - \widetilde{p}, \widetilde{x}_2 - \widetilde{p} \rangle$ we use that $\angle \widetilde{x}_1 \widetilde{p} \widetilde{x}_2 = \angle uvw + \angle v \widetilde{x}_2 \widetilde{p}$, that the sine rule says $\|\widetilde{x}_2 - \widetilde{p}\| \sin(\angle \widetilde{x}_2 v \widetilde{p}) = \|v - \widetilde{p}\| \sin \angle uvw$, and that $\cos \angle v \widetilde{x}_2 \widetilde{p} > 0$, $\cos \angle uvw < 0$ and $-\sin^2 \angle uvw \leq -\sin \angle uvw$. Together these imply that

$$\begin{aligned}
\langle \widetilde{x}_1 - \widetilde{p}, \widetilde{x}_2 - \widetilde{p} \rangle &= \|\widetilde{x}_1 - \widetilde{p}\| \|\widetilde{x}_2 - \widetilde{p}\| \cos(\angle uvw + \angle v \widetilde{x}_2 \widetilde{p}) \\
&= \|\widetilde{x}_1 - \widetilde{p}\| \|\widetilde{x}_2 - \widetilde{p}\| \cos \angle uvw \cos(\angle v \widetilde{x}_2 \widetilde{p}) - \sin^2 \angle uvw \|v - \widetilde{p}\| \|\widetilde{x}_1 - \widetilde{p}\| \\
&\geq (R+\varepsilon)(R+3\varepsilon) \cos \angle uvw - \sin \angle uvw \|v - \widetilde{p}\| (R+\varepsilon).
\end{aligned}$$

From the assumptions in this scenario that $\|v - \widetilde{p}\| \leq 4\varepsilon$, we can substitute into (1.3) to get

$$\begin{aligned}
&\langle x_1 - p, x_2 - p\rangle \\
&\geq (R+\varepsilon)(R+3\varepsilon)\cos\angle uvw - 4\varepsilon(R+\varepsilon)\sin\angle uvw - R\varepsilon(2+\sin\angle uvw) - (2+\sin\angle uvw)\varepsilon^2 \\
&= (R+\varepsilon)\sqrt{R^2+6R\varepsilon+34\varepsilon^2}\sin\left(\angle uvw + \arctan\left(\frac{R+3\varepsilon}{5\varepsilon}\right)\right) - 2\varepsilon R - 2\varepsilon^2.
\end{aligned}$$

From our assumptions on $\angle uvw$

$$\sin\left(\angle uvw + \arctan\left(\frac{R+3\varepsilon}{5\varepsilon}\right)\right) > -\frac{R^2-4R\varepsilon+\varepsilon^2}{(R+\varepsilon)\sqrt{R^2+6R\varepsilon+34\varepsilon^2}}.$$

Thus we conclude

$$\begin{aligned}
&\langle x_1 - p, x_2 - p\rangle \\
&> (R+\varepsilon)\sqrt{R^2+6R\varepsilon+34\varepsilon^2}\left(-\frac{R^2-4R\varepsilon-9\varepsilon^2}{(R+\varepsilon)\sqrt{R^2+6R\varepsilon+34\varepsilon^2}}\right) - 2\varepsilon R - 2\varepsilon^2 \\
&= -R^2 + 2R\varepsilon + 7\varepsilon^2.
\end{aligned}$$

Now assume we are in scenario 2: $\|v-p\| < (R-\varepsilon)/2$ and $\angle uvw \leq \pi/2$.

To prove the claim in this scenario, we will need to further split into two cases;

(i) $\angle \widetilde{x_1}\widetilde{p}\widetilde{x_2} \leq \pi/2$, and
(ii) $\angle \widetilde{x_1}\widetilde{p}\widetilde{x_2} > \pi/2$.

In case (i) we have $\langle \widetilde{x_1} - \widetilde{p}, \widetilde{x_2} - \widetilde{p}\rangle \geq 0$ and thus

$$\langle x_1 - p, x_2 - p\rangle \geq -3R\varepsilon - 3\varepsilon^3 > -R^2 + 2R\varepsilon + 7\varepsilon^2$$

as $R > 12\varepsilon$.

In case (ii), thinking of the inner product in terms of the projection of vector $\widetilde{x_2} - \widetilde{p}$ onto $\widetilde{x_1} - \widetilde{p}$ we get

$$
\begin{aligned}
\langle x_1 - p, x_2 - p\rangle &\geq -\|\widetilde{x_1} - \widetilde{p}\|\|v - \widetilde{p}\| - 3R\varepsilon - 3\varepsilon^3 \\
&\geq -(R+\varepsilon)(R-\varepsilon)/2 - 3R\varepsilon - 3\varepsilon^3 \\
&= -R^2/2 - 3R\varepsilon - 5\varepsilon^2/2 \\
&> -R^2 + 2R\varepsilon + 7\varepsilon^2,
\end{aligned}
$$

where in the final inequality we use that $R > 12\varepsilon$. □

To find sufficient conditions for detecting when a sample p is near a vertex, we want each edge adjacent to that vertex to correspond to at least one distinct cluster of points in the spherical shell around p. To avoid the clusters around separate edges merging, we assume a lower bound on the angle between the edges as part of our assumptions on the geometric embedding. The following lemma motivates this choice of lower bound.

Lemma 1.1.3. *Let $u, v, w \in \mathbb{R}^n$, $D > \varepsilon > 0$, and let $x_1, x_2 \in \mathbb{R}^n$ satisfy*

1. $d(x_1, \overline{uv}), d(x_2, \overline{uw}) < \varepsilon$, *and*
2. $\|x_1 - v\|, \|x_2 - v\| > D$.

If

$$\angle uvw > \arccos\left(\frac{2D^2 - 9\varepsilon^2}{2D^2}\right) + 2\arcsin\left(\frac{\varepsilon}{D}\right)$$

then $\|x_1 - x_2\| > 3\varepsilon$.

Proof. The distance between x_1 and x_2 is minimised when

$$\|v - x_1\| = D = \|v - x_2\|.$$

Furthermore we can observe that $\angle uvx_1 = \arcsin\left(\frac{d(x_1,\overline{uv})}{\|x_1 - v\|}\right) \leq \arcsin(\varepsilon/D)$. Similarly $\angle uvx_1 \leq \arcsin(\varepsilon/D)$. This implies

$$\angle x_1 v x_2 \geq \angle uvw - \angle uvx_1 - \angle wvx_2 \geq \alpha - 2\arcsin(\varepsilon/D).$$

Combining we conclude

$$\begin{aligned}\|x_1 - x_2\|^2 &\geq \|v - x_1\|^2 + \|v - x_2\|^2 - \|v - x_1\|\|v - x_2\| \cos \angle x_1 v x_2 \\ &\geq 2D^2 - 2D^2 \cos(\alpha - 2\arcsin(\varepsilon/D)) \\ &\geq (3\varepsilon)^2.\end{aligned}$$

□

1.2. Determining Local Structure

Given an ε-sample P of an embedded graph $|G|$, we want to recover the abstract graph G by approximating the local structure of $|G|$ at each sample $p \in P$. When approximating the local structure at a sample p, we regularly consider the graph on a set of points, with edges (p, q) if $\|p - q\| \leq r$, for some fixed $r \in \mathbb{R}$.

Definition 1.2.1. *Let $P \subset \mathbb{R}^N$ be a finite collection of points, and fix $r > 0$. The* graph at threshold r on P, *$\mathfrak{G}_r(P)$, is the graph with vertices $p \in P$, and edges (p, q) if $\|p - q\| \leq r$.*

For each $p \in P$, we will consider two graphs on points close to p: the first approximates the connectedness of $|G|$ intersected with a ball around p, the second consists of points in a spherical shell around p. We call this pair of graphs the *(R, ε)-local structure of P at p.*

Definition 1.2.2 ((R, ε)-local structure)**.** *Let $P \subset \mathbb{R}^n$ be an ε-sample of an embedded graph $|G|$ and fix $R > 12\varepsilon$. The (R, ε)-local structure of P at a sample $p \in P$ is the pair*

$$\left(\mathfrak{G}_{3\varepsilon}(P \cap B_{R+\varepsilon}(p)), \mathfrak{G}_{3\varepsilon}(P \cap S_{R-\varepsilon}^{R+\varepsilon}(p))\right).$$

We want to use the (R, ε)-local structure to approximate $|G| \cap B_R(p)$ for each $p \in P$, and use this to learn the structure of $|G|$. We will classify samples as being near a vertex or not near a vertex by their (R, ε)-local structure.

We now formalise what the (R, ε)-local structure is for points $p \in P$ not near any vertex $v \in |G|$. That is, points which have (R, ε)-local structure of an edge.

Definition 1.2.3 (Local structure of an edge)**.** *Let P be an ε-sample of a linearly embedded graph $|G|$. A point $p \in P$ has the (R, ε)*-local structure of an edge *if either of the following hold:*

1. *$\mathfrak{G}_{3\varepsilon}(P \cap B_{R+\varepsilon}(p))$ is disconnected,*
2. *$\mathfrak{G}_{3\varepsilon}(P \cap B_{R+\varepsilon}(p))$ is connected, $\mathfrak{G}_{3\varepsilon}(P \cap S^{R+\varepsilon}_{R-\varepsilon}(p))$ has two connected components c_1, c_2 with average points q_1 and q_2, and*

$$\langle q_1 - p, q_2 - p\rangle \leq -R^2 + 2R\varepsilon + 7\varepsilon^2.$$

We now define the (R, ε)*-local structure of a vertex.*

Definition 1.2.4 (Local structure of a vertex)**.** *Let P be an ε-sample of a linearly embedded graph $|G|$. A point $p \in P$ has the (R, ε)*-local structure of a vertex *if either of the following hold:*

1. *$\mathfrak{G}_{3\varepsilon}(P \cap B_{R+\varepsilon}(p))$ is connected, and the number of connected components in $\mathfrak{G}_{3\varepsilon}(P \cap S^{R+\varepsilon}_{R-\varepsilon}(p))$ is not 2,*
2. *$\mathfrak{G}_{3\varepsilon}(P \cap B_{R+\varepsilon}(p))$ is connected, $\mathfrak{G}_{3\varepsilon}(P \cap S^{R+\varepsilon}_{R-\varepsilon}(p))$ has two connected components c_1, c_2 with average points $q_1 = \frac{1}{|c_1|}\Sigma_{p\in c_1} p$ and $q_2 = \frac{1}{|c_1|}\Sigma_{p\in c_2} p$, and*

$$\langle q_1 - p, q_2 - p\rangle > -R^2 + 2R\varepsilon + 7\varepsilon^2.$$

Next, we formally define P_0 and P_1.

Definition 1.2.5 (P_0 and P_1)**.** *Given an ε-sample P of a linearly embedded graph $|G| \subset \mathbb{R}^n$, we define the partitioning sets P_0 and P_1 as follows:*

$$P_0 = \{p \in P \mid \ p \textit{ has the } (R, \varepsilon)\textit{-local structure of a vertex.}\}$$

$$P_1 = \{p \in P \mid \ p \textit{ has the } (R, \varepsilon)\textit{-local structure of an edge.}\}$$

Remark 1.2.6. Note that a sample $p \in P$ has either (R, ε)-local structure of a vertex (R, ε)-local structure of an edge. Hence, the partitioning defined in Definition 1.2.5 is disjoint.

As we use the connected components of $\mathfrak{G}_{3\varepsilon}(P \cap B_{R+\varepsilon}(p))$ and $\mathfrak{G}_{3\varepsilon}(P \cap S^{R+\varepsilon}_{R-\varepsilon}(p))$ in the definition of the (R, ε)-local structure of p, we require some assumptions on $|G|$ to ensure that we correctly identify when points are near vertices or not. To ensure $\mathfrak{G}_{3\varepsilon}(P \cap B_{R+\varepsilon}(p))$ is not disconnected for points p near some vertex, we assume that the distance between a vertex v and any edge $\overline{uw}$, $u, w \neq v$, is bounded below $d(v, \overline{wv}) > R + \frac{R}{2} + 2\varepsilon$. To ensure that there are samples near edges which are not near any vertex, we additionally assume that for every pair of vertices u, v, $\|u - v\| > \frac{9R}{2} + 6\varepsilon$.

We also place lower and upper bounds on the angles between edges. For ease of notation, we will define two functions for these bounds.

Definition 1.2.7. *Given $R > 12\varepsilon$, we set*

$$\Psi(R,\varepsilon) := \pi - \arctan\left(\frac{R+3\varepsilon}{6\varepsilon}\right) + \arcsin\left(\frac{R^2 - 4R\varepsilon - 9\varepsilon^2}{(R+\varepsilon)\sqrt{R^2+6R\varepsilon+34\varepsilon^2}}\right),$$

$$\Phi(R,\varepsilon) := \arccos\left(\frac{(R-\varepsilon)^2 - 18\varepsilon^2}{(R-\varepsilon)^2}\right) + 2\arcsin\left(\frac{2\varepsilon}{(R-\varepsilon)}\right).$$

To improve intuition of these functions, Figures 1.2 and 1.3 provide graphs of them. Note they are effectively a function of $\frac{R}{\varepsilon}$ as they are invariant to scaling both R and ε by the same amount.

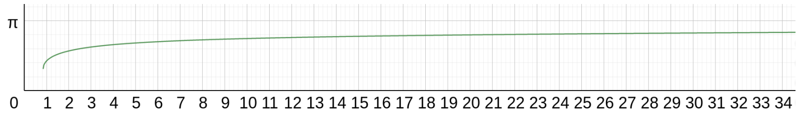


FIGURE 1.2. Graph of $\Psi\left(\frac{R}{\varepsilon}, 1\right)$.

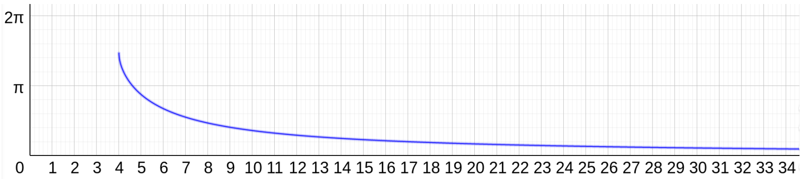


FIGURE 1.3. Graph of $\Phi\left(\frac{R}{\varepsilon}, 1\right)$.

Henceforth, we assume that all embedded graphs $|G|$ satisfy the following assumptions.

Assumption 1. *Fix $R \geq 12\varepsilon > 0$. We restrict to embedded graphs $|G| = (G, \phi_G)$ satisfying the following.*

1. *For all vertices u, v, $\|u - v\| > \frac{9R}{2} + 6\varepsilon$.*
2. *For a vertex v and an edge $\overline{uw}$, with $u, w \neq v$, $d(v, \overline{uw}) > \frac{3R}{2} + 4\varepsilon$.*
3. *For any pair of edges $\overline{uv}, \overline{xy}$ with no common vertex, $d(\overline{uv}, \overline{xy}) > 5\varepsilon$.*
4. *For all pairs of edges $\overline{uv}, \overline{wv}$, $\angle uvw \geq \Phi(R, \varepsilon)$.*
5. *For all degree 2 vertices v with edges $\overline{uv}, \overline{wv}$, $\angle uvw \leq \Psi(R, \varepsilon)$.*

The propositions in this section are used to show that the clusters in P_0 and P_1 correspond bijectively with the vertices and edges of $|G|$. The first proposition shows that for all samples p near a vertex v with deg$(v) \neq 2$, p has the (R, ε)-local structure of a vertex. The second and third prove that samples near degree 2 vertices also have the (R, ε)-local structure of a vertex. The final proposition shows that all samples p not near any vertex have the (R, ε)-local structure of an edge.

Proposition 1.2.8. *Let v be a vertex of $|G| \subset \mathbb{R}^n$ with deg$(v) \neq 2$, and let P be an ε-sample of $|G|$. Then for all $p \in P$ with $\|p - v\| \leq \frac{R-\varepsilon}{2}$, p has the (R, ε)-local structure of a vertex.*

Proof. We begin by considering $\mathbf{deg(v) = 0}$. By Assumption 1 (1), $\|p - v\| \leq \varepsilon$, and for all $q \in P \cap B(p, R+\varepsilon)$, $\|q - v\| \leq \varepsilon$. Thus $\mathfrak{G}_{3\varepsilon}\left(P \cap B_{R+\varepsilon}(p)\right)$ is connected. Similarly, $P \cap S^{R+\varepsilon}_{R-\varepsilon}(p) = \emptyset$, and p has the (R, ε)-local structure of a vertex.

Next, assume $\mathbf{deg(v) = 1}$. For the edge $\overline{uv}$, let $t_0, t_1, \ldots, t_m$ be consecutive points along $\overline{uv}$ with $\|t_0 - v\|$, $\|t_{i+1} - t_i\| \leq \varepsilon$ and $\|p - t_m\| = R + \varepsilon$. Then, there must be $z_0, z_1, \ldots, z_m \in P$ with $\|t_i - z_i\| \leq \varepsilon$. Note, these z_i may not be unique. Since $\|z_i - z_{i+1}\| \leq 3\varepsilon$, and every sample in $P \cap B_{R+\varepsilon}(p)$ is within 3ε of some z_i, $\mathfrak{G}_{3\varepsilon}\left(P \cap B_{R+\varepsilon}(p)\right)$ is connected.

If the number of clusters in $\mathfrak{G}_{3\varepsilon}(P \cap S^{R+\varepsilon}_{R-\varepsilon}(p))$ is not 2, then p has the (R, ε)-local structure of a vertex. Thus suppose that there are 2 connected components. We will show that inner product condition between their averages will declare that p has the (R, ε)-local structure of a vertex.

Let $x_1, x_2 \in P \cap S^{R+\varepsilon}_{R-\varepsilon}(p)$ be samples in the two connected components c_1 and c_2. Observe that both x_1 and x_2 are within ε of the line segment $\overline{uv}$.

As $\|p-v\|\leq \frac{R-\varepsilon}{2}$, and x_1, x_2 are within ε of the same edge $\overline{uv}$, x_1 and x_2 are contained on the same side of hyper-plane H through p perpendicular to $\overline{vu}$.

We can observe that $\angle x_1 p x_2 \leq 2\arccos\left(\frac{2\varepsilon}{R-\varepsilon}\right) < \pi/2$, and thus

$$\langle x_1 - p, x_2 - p\rangle > 0 > -R^2 + 2R\varepsilon + 7\varepsilon^2.$$

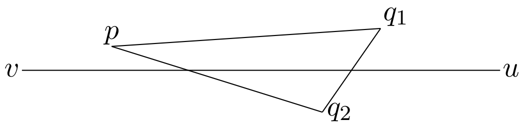


FIGURE 1.4. Both q_1 and q_2 are in the same half-space generated by the hyper-plane through p perpendicular to $\overline{uv}$.

As this holds for all $x_1 \in c_1, x_2 \in c_2$, it also holds for the averages q_1 and q_2. Thus p has the (R,ε)-local structure of a vertex.

Finally, assume $\mathbf{deg(v)} \geq \mathbf{3}$. From analogous arguments as in the degree 1 case we know that $\mathfrak{G}_{3\varepsilon}\left(P \cap B_{R+\varepsilon}(p)\right)$ is connected.

Now consider $S^{R+\varepsilon}_{R-\varepsilon}(p)$. For each edge $\overline{uv}$, there is a sample $x_{\overline{uv}} \in S^{R+\varepsilon}_{R-\varepsilon}(p)$. To show there are at least 3 connected components in $\mathfrak{G}_{3\varepsilon}(P \cap S^{R+\varepsilon}_{R-\varepsilon}(p))$, we need only check that samples from different edges cannot merge to be in the same connected component in $\mathfrak{G}_{3\varepsilon}(P \cap S^{R+\varepsilon}_{R-\varepsilon}(p))$. By way of contradiction suppose there were edges $\overline{uv}$ and $\overline{wv}$ and samples $x_u, x_v \in P \cap S^{R+\varepsilon}_{R-\varepsilon}(p)$ within ε of $\overline{uv}$ and $\overline{wv}$ respectively such that $\|x_u - x_w\|\leq 3\varepsilon$. As $\|p-v\|\leq (R-\varepsilon)/2$ and $\|p-x_u\|, \|p-x_v\|\geq R-\varepsilon$ we know $\|v-x_u\|, \|v-x_w\|> (R-\varepsilon)/2$. This contradicts Lemma 1.1.3 as this implies that $\|x_u - x_v\|> 3\varepsilon$.

We conclude that $\mathfrak{G}_{3\varepsilon}(P \cap S^{R+\varepsilon}_{R-\varepsilon}(p))$ has at least as many connected components as the degree of v. Thus, p has the (R,ε)-local structure of a vertex. □

Proposition 1.2.9. *Let v be a vertex of $|G|\subset \mathbb{R}^n$ with deg$(v) = 2$, with edges $\overline{uv}, \overline{wv}$. Let P be an ε-sample of $|G|$. If $\angle uwv > \frac{\pi}{2}$, then for all $p \in P$ with $\|p-v\|\leq 4\varepsilon$, p has the (R,ε)-local structure of a vertex.*

Proof. As in the proof of Proposition 1.2.8, $\mathfrak{G}_{2\varepsilon}(P \cap B_{R+\varepsilon}(p))$ is connected.

For both edges $\overline{uv}, \overline{wv}$ there is at least one sample in $S^{R+\varepsilon}_{R-\varepsilon}(P)$, say $q_{\overline{uv}}$ and $q_{\overline{wv}}$. By Lemma 1.1.3, for all q' in $S^{R+\varepsilon}_{R_\varepsilon} \cap P$, if $d(q', q_{\overline{wv}}) \leq 3\varepsilon$, then $\|q' - q\|> 3\varepsilon$. Hence, each edge contributes at least 1 connected component to $\mathfrak{G}_{3\varepsilon}(P \cap S^{R+\varepsilon}_{R-\varepsilon}(p))$.

If there are more than 2, then p has the (R, ε)-local structure of a vertex. We now assume there are 2 connected components c_1, c_2 (one per edge) in $\mathfrak{G}_{3\varepsilon}(P \cap S^{R+\varepsilon}_{R-\varepsilon}(p))$. Lemma 1.1.2 gives

$$\langle q_1 - p, q_2 - p\rangle > -R^2 + 2R\varepsilon + 7\varepsilon^2,$$

and p has the (R, ε)-local structure of a vertex. □

Proposition 1.2.10. *Let v be a vertex of $|G|\subset \mathbb{R}^n$ with deg$(v) = 2$, with edges $\overline{uv}, \overline{wv}$. Let P be an ε-sample of $|G|$. If $\angle uvw \leq \frac{\pi}{2}$, then for all $p \in P$ with $\|p - v\|\leq \frac{R-\varepsilon}{2}$, p has the (R, ε)-local structure of a vertex.*

Proof. As in the proof of Proposition 1.2.8, $\mathfrak{G}_{2\varepsilon}(P \cap B_{R+\varepsilon}(p))$ is connected.

For both edges $\overline{uv}, \overline{wv}$ there is at least one sample in $S^{R+\varepsilon}_{R-\varepsilon}(P)$, say $q_{\overline{uv}}$ and $q_{\overline{wv}}$. By Lemma 1.1.3, for all q' in $S^{R+\varepsilon}_{R_\varepsilon} \cap P$, if $\|q' - q_{\overline{wv}}\|\leq 3\varepsilon$, then $\|q' - q\|> 3\varepsilon$. Hence, each edge contributes at least 1 connected component to $\mathfrak{G}_{3\varepsilon}(P \cap S^{R+\varepsilon}_{R-\varepsilon}(p))$.

If there are more than 2, then p has the (R, ε)-local structure of a vertex. We now assume there are 2 connected components c_1, c_2 (one per edge) in $\mathfrak{G}_{3\varepsilon}(P \cap S^{R+\varepsilon}_{R-\varepsilon}(p))$.

Let x_1 and x_2 be points in c_1 and c_2. Without loss of generality, we have $d(x_1, \overline{uv}), d(x_2, \overline{wv}) \leq \varepsilon$.

From Lemma 1.1.2 we know that $\langle x_1 - p, x_2 - p\rangle < -R^2 + 2R\varepsilon + 7\varepsilon^2$. Since this inequality holds for all pairs x_1, x_2 in the connected components c_1 and c_2 we know it also holds for the averages q_1 and q_2. Thus we conclude p has the (R, ε)-local structure of a vertex.

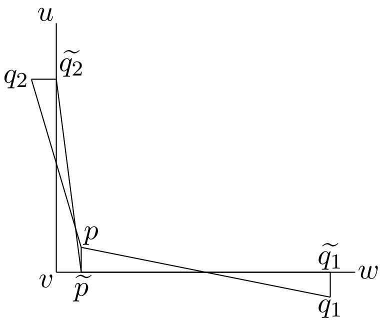


FIGURE 1.5

FIGURE 1.6. The case where $\angle uvw \leq \frac{\pi}{2}$.

□

Proposition 1.2.11. *Let $p \in P$ be a sample with $\|p - v\| > \frac{3R+\varepsilon}{2}$ for all vertices $v \in |G|$. Then p has the (R, ε)-local structure of an edge.*

Proof. We begin by showing that if there is a sample $q \in S_{R-\varepsilon}^{R+\varepsilon}(p) \cap P$ with $d(q, \overline{uv}) > \varepsilon$, then $\mathfrak{G}_{3\varepsilon}(B_{R+\varepsilon}(p) \cap P)$ is disconnected. To prove this suppose not. Then there exists $x, y \in B_{R+\varepsilon}(p) \cap P$ such that $d(x, \overline{uv}) < \varepsilon$, $d(y, \overline{uv}) > \varepsilon$ and yet $\|x - y\| < 3\varepsilon$.

This splits into two cases:

(i) $d(y, \overline{wv}) \leq \varepsilon$ for some vertex $w \neq u$ (noting that this case covers an edge $\overline{wu}$ as well),

(ii) $d(y, \overline{wz}) \leq \varepsilon$ for vertices $w, z \neq u, v$.

For case (i), first observe that $\|x - v\|, \|y - v\| > \frac{R-\varepsilon}{2}$. We then get a contradiction via Lemma 1.1.3 (with $D = \frac{R-\varepsilon}{2}$) using Assumption 1 (4).

For case (ii) recall that Assumption 1 (3) implies $d(\overline{uv}, \overline{wz}) > 5\varepsilon$. However $d(\overline{uv}, \overline{wz}) < d(\overline{uv}, x) + \|x - y\| + d(y, \overline{wz}) \leq 5\varepsilon$ which is a contradiction.

We thus conclude that if there is some $q \in S_{R-\varepsilon}^{R+\varepsilon}(p) \cap P$ with $d(q, \overline{uv}) > \varepsilon$ then $\mathfrak{G}_{3\varepsilon}(P \cap B_{R+\varepsilon}(p))$ is disconnected and p has the (R, ε)-local structure of an edge.

We can now assume that $\mathfrak{G}_{3\varepsilon}(P \cap B_{R+\varepsilon}(p))$ is connected, and for all $q \in P \cap B_{R+\varepsilon}(p)$, $d(q, \overline{uv}) \leq \varepsilon$. We need to show that there are two clusters of samples in $S_{R-\varepsilon}^{R+\varepsilon}(p)$. Let $n \in \overline{uv}$ satisfy $\|p - n\| = R$, and assume that n

and q are on the same side of the hyper-plane H through p perpendicular to $\overline{uv}$. Now let $\widetilde{p}, \widetilde{q}$ be the projections of p and q respectively to $\overline{uv}$.

We will split the analysis into the cases where $\|\tilde{p} - \tilde{q}\| \leq \|\tilde{p} - n\|$ and where $\|\tilde{p} - \tilde{q}\| > \|\tilde{p} - n\|$.

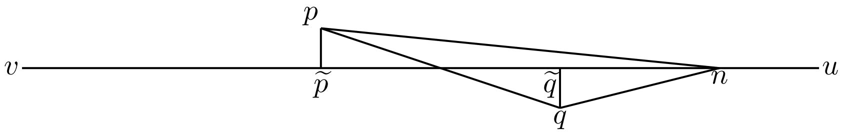


FIGURE 1.7. The case where $\|\widetilde{p} - \widetilde{q}\| < \|\widetilde{p} - n\|$.

Consider $\|\widetilde{p} - \widetilde{q}\| \leq \|\widetilde{p} - n\|$, as in Figure 1.7. Note that $\|\tilde{p} - n\| \leq R$ and $\|\tilde{p} - \tilde{q}\| \geq \sqrt{(R-\varepsilon)^2 - (2\varepsilon)^2}$ which implies

$$\begin{aligned} \|q - n\|^2 &= \|q - \widetilde{q}\|^2 + \left(\|\tilde{p} - n\| - \|\widetilde{p} - \widetilde{q}\|^2\right) \\ &\leq \varepsilon^2 + \left(R - \sqrt{(R-\varepsilon)^2 - 4\varepsilon^2}\right)^2 . \end{aligned} \tag{1.4}$$

Now consider $\|\widetilde{p} - n\| < \|\widetilde{p} - \widetilde{q}\|$, such as in Figure 1.8. Here we use the bounds $\|\tilde{p} - n\| \geq \sqrt{R^2 - \varepsilon^2}$ and $\|\tilde{p} - \tilde{q}\| \leq R + \varepsilon$ to say

$$\begin{aligned} \|q - n\|^2 &= \|q - \widetilde{q}\|^2 + (\|\widetilde{p} - \widetilde{q}\| - \|\widetilde{p} - n\|)^2 \\ &\leq \varepsilon^2 + \left(\sqrt{(R+\varepsilon)^2} - \sqrt{R^2 - \varepsilon^2}\right)^2 . \end{aligned} \tag{1.5}$$

Algebraic manipulation shows that both (1.4) and (1.5) are bounded from above by $4\varepsilon^2$ whenever $R > 12\varepsilon$.

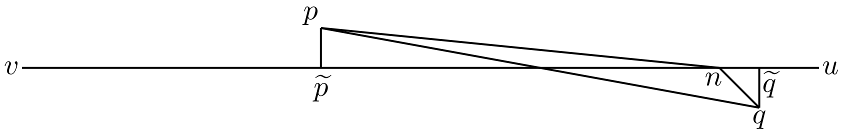


FIGURE 1.8. The case where $\|\widetilde{p} - \widetilde{q}\| > \|\widetilde{p} - n\|$.

Thus, for all q on the same side of H as n with $\|p - q\| \leq R$, we have $\|q - n\| \leq 2\varepsilon$.

As $n \in \overline{uv}$, there is a sample $q_n \in P$ with $\|n - q_n\| \leq \varepsilon$. Importantly since $B_\varepsilon(n) \subset S_{R-\varepsilon}^{R+\varepsilon}(p)$ we can say that q_n connects to all the $P \cap S_{R-\varepsilon}^{R+\varepsilon}(p)$ on the same side of H within $\mathfrak{G}_{3\varepsilon}\left(P \cap S_{R-\varepsilon}^{R+\varepsilon}(p)\right)$.

Thus, on each side of H, we have a single cluster of points, which are connected at 3ε. Thus, $\mathfrak{G}_{3\varepsilon}\left(P \cap S_{R-\varepsilon}^{R+\varepsilon}(p)\right)$ has two connected components. Then, Lemma 1.1.1 implies that p has the (R, ε)-local structure of an edge.

□

1.3. Algorithm and Its Correctness

In this section, we present the algorithm from Skyler, and prove that given P an ε-sample of an embedded graph $|G| = (G, \phi_G)$ satisfying Assumptions 1, the algorithm returns an isomorphic graph structure. The algorithm partitions P into P_0 and P_1, such that for each $p \in P_0$, p has the (R, ε)-local structure of a vertex, and for each $p \in P_1$, p has the (R, ε)-local structure of an edge. We then detect the number of vertices, the number of edges and the boundary operator. To obtain P_0 and P_1, we use the function $\Delta_{R,\varepsilon} : P \to \{0, 1\}$, (Algorithm 1), such that if p has (R, ε)-local structure of a vertex $\Delta_{R,\varepsilon}(p) = 0$ and if p (R, ε)-local structure of an edge, $\Delta_{R,\varepsilon}(p) = 1$. Then, $P_0 = \Delta_{R,\varepsilon}^{-1}(0)$ and $P_1 = \Delta_{R,\varepsilon}^{-1}(1)$.

For each vertex $v \in |G|$, if $\deg(v) \neq 2$, Proposition 1.2.8 implies that for all $p \in P$ with $\|p - v\| \leq \frac{R}{2}$, $\Delta_{R,\varepsilon}(p) = 0$, while if $\deg(v) = 2$, Propositions 1.2.9 and 1.2.10 imply that $\Delta_{R,\varepsilon}(p) = 0$, and Proposition 1.2.11 implies that if $\|p - v\| > \frac{3R}{2} + 2\varepsilon$, $\Delta_{R,\varepsilon}(p) = 1$.

Lemma 1.3.1. *Let $x \in P_0$ and $\|x - v\| < \frac{3R}{2} + \varepsilon$ for vertex v. Then $y \in P_0$ is in the same connected component as x in $\mathfrak{G}_{\frac{3R}{2}+2\varepsilon}(P_0)$ if and only if $\|y - v\| < \frac{3R}{2} + \varepsilon$.*

Proof. By Proposition 1.2.11 $P_0 \subset P \cap \left\{\bigcup_{v \in V} B_{\frac{3R}{2}+\varepsilon}(v)\right\}$. Our embedding assumptions require that for vertices $v \neq v'$ we have $\|v - v'\| > \frac{9R}{2} + 3\varepsilon$ and hence no points in $P \cap B_{\frac{3R}{2}+\varepsilon}(v')$ are within $\frac{3R}{2} + \varepsilon$ of those in $B_{\frac{3R}{2}+\varepsilon}(v')$. This means they can not be connected in $\mathfrak{G}_{\frac{3R}{2}+2\varepsilon}(P_0)$. This implies that the entire connected component containing x must lie in $B_{\frac{3R}{2}+\varepsilon}(v)$. If $\|y - v\| > \frac{3R}{2} + \varepsilon$ then it cannot be in the same connected component as x.

We finally wish to show that $\|y-v\|< \frac{3R}{2}+\varepsilon$ implies that x and y are in the same connected component. Choose vertices u_y and u_x such that $d(y, \overline{u_y v}) < \varepsilon$ and $d(x, \overline{u_x v}) < \varepsilon$. Now let $z_y \in \overline{uv}$ be the point 3ε from v. We analogously define z_x. As P is an ε-sample of $|G|$ we have samples p_y and p_x such that $\|p_y - z_y\|< \varepsilon$ and $\|p_x - z_x\|< \varepsilon$. Note that $p_y, p_x \in P \cap B_{4\varepsilon}(v)$ and hence by Propositions 1.2.8 and 1.2.9 we know that $p_y, p_x \in P_0$. By construction $\|y-p_y\|$, $\|p_y - p_x\|$ and $\|p_x - x\|$ are all less that $\frac{3R}{2}+\varepsilon$ and hence y and x are in the same connected component in $\mathfrak{G}_{\frac{3R}{2}+2\varepsilon}(P_0)$. □

The above lemma shows the correspondence between vertices in G and connected components in $\mathfrak{G}_{\frac{3R}{2}+2\varepsilon}(P_0)$. Unfortunately the situation is less clean for the connected components of $\mathfrak{G}_{3\varepsilon}(P_1)$. Around each vertex v there is a 'grey area', in which samples p can be placed in either P_0 or P_1. Due to the size of this spherical shell, it is possible to obtain connected components in $\mathfrak{G}_{3\varepsilon}(P_1)$ which contain points only within such a grey area. We devote the next few results to characterising the connected components of $\mathfrak{G}_{3\varepsilon}(P_1)$. We first show that every connected component of $\mathfrak{G}_{3\varepsilon}(P_1)$ is close to only one edge.

Proposition 1.3.2. *Let $[x]$ be a connected component of $\mathfrak{G}_{3\varepsilon}(P_1)$. Then there exists an edge $\overline{uv}$ such that $d(y, \overline{uv}) < \varepsilon$ for all $y \in [x]$.*

Proof. Since every sample in P is within ε of some edge it is sufficient to show that if $p, q \in P_1$ with $d(p, \overline{uv}) \leq \varepsilon$ and $\|p-q\|\leq 3\varepsilon$ then $d(q, \overline{uv}) < \varepsilon$.

As $p \in P_1$, Propositions 1.2.8, 1.2.9 and 1.2.10 imply

1. for all vertices $w \in |G|$ with $\deg(w) \neq 2$, $\|p-w\|> \frac{R-\varepsilon}{2}$,
2. for all vertices w with $\deg(w) = 2$, $\|p-w\|\geq 4\varepsilon$.

Without loss of generality, assume $\|p-v\|\leq \|p-u\|$. By Assumptions 1 (3) for all edges $\overline{xy}$ with x, y distinct from u, v, $d(\overline{uv}, \overline{xy}) > 5\varepsilon$. Hence, $d(p, \overline{xy}) > 4\varepsilon$, and for any sample q with $d(q, \overline{xy}) \leq \varepsilon$, $\|p-q\|> 3\varepsilon$. If $\deg(v) \neq 2$, then $\|p-v\|> \frac{R-\varepsilon}{2}$, and as $|G|$ satisfies Assumptions 1 (4), Lemma 1.1.3 implies $\|p-q\|> 3\varepsilon$ for all $q \in P_1$ with $d(q, \overline{uv}) > \varepsilon$.

Now assume that $\deg(v) = 2$, and consider another edge $\overline{wv}$. For $\Phi(R, \varepsilon) \leq \angle uvw < \frac{\pi}{2}$ we can apply Lemma 1.1.3 with $D = \frac{R-\varepsilon}{2}$ to see that for all $q \in P_1$ with $d(q, \overline{wv}) \leq \varepsilon$, $\|q, -p\|> 3\varepsilon$. For $\frac{\pi}{2} \leq \angle uvw \leq$

$\Psi(R,\varepsilon)$, we apply Lemma 1.1.3 with $D = 4\varepsilon$ and observe that $\pi/2 > \arccos(23/32) + 2\arcsin(1/4)$ to conclude that $d(q, \overline{wv}) \leq \varepsilon$, $\|q - p\| > 3\varepsilon$. □

There can be multiple connected component in $\mathfrak{G}_{3\varepsilon}(P_1)$ near the same edge. However there will only be one which contains a sample near the midpoint of the edge. We wish to treat these differently and so we will give them a name.

Definition 1.3.3. *We say that the connected component of $\mathfrak{G}_{3\varepsilon}(P_1)$* spans *the edge $\overline{uv}$ if it contains a point within ε of the midpoint of $\overline{uv}$. Without reference to the specific edge $\overline{uv}$ we say that the component is* spanning.

Proposition 1.3.4. *Let $\overline{uv}$ be an edge in G. There exists a unique connected component $A_{\overline{uv}}$ which spans $\overline{uv}$. $A_{\overline{uv}}$ contains samples in both $B_{\frac{3R+5\varepsilon}{2}}(u)$ and $B_{\frac{3R+5\varepsilon}{2}}(v)$.*

If $[x] \neq A_{\overline{uv}}$ is a connected component in $\mathfrak{G}_{3\varepsilon}(P_1)$ within ε of $\overline{uv}$ then either $[x] \subset B_{\frac{3R+\varepsilon}{2}}(u)$ or $[x] \subset B_{\frac{3R+\varepsilon}{2}}(v)$.

Proof. Let m denote the midpoint of $\overline{uv}$.

Let $t_0, t_1, \ldots t_{2M}$ be consecutive points along $\overline{uv}$ with $\|t_i - t_{i+1}\| < \varepsilon$, $\|t_0 - u\| = \frac{3R+3\varepsilon}{2}$, $t_M = m$, and $\|t_{2M} - v\| = \frac{3R+3\varepsilon}{2}$. There must be $z_0, z_1 z_2, \ldots z_M \in P$ such that $\|t_i - z_i\| < \varepsilon$. Observe that $\|z_i - u\| > \frac{3R+\varepsilon}{2}$ and $\|z_i - v\| > \frac{3R+\varepsilon}{2}$ and so by Proposition 1.2.11 all the z_i are in P_1. Since $\|z_i - z_{i+1}\| < 3\varepsilon$ we know that all the z_i lie in the same connected component of $\mathfrak{G}_{3\varepsilon}(P_1)$ which spans $\overline{uv}$ as z_M is within ε of m.

To see this connected component is unique we need only observe that any pair of samples in P_1 both within ε of m are within 3ε of each other and hence lie in the same connected component. Denote this unique connected component by $A_{\overline{uv}}$.

Observe that $\|u - z_0\| < \frac{3R+3\varepsilon}{2} + \varepsilon$ and $\|v - z_{2M}\| < \frac{3R+3\varepsilon}{2} + \varepsilon$.

Now suppose that $[x] \neq A_{\overline{uv}}$ is a connected component in $\mathfrak{G}_{3\varepsilon}(P_1)$ within ε of $\overline{uv}$. Since $[x] \neq A_{\overline{uv}}$, we have $d([x], t_i) > 2\varepsilon$ for all i and hence

$$[x] \subset B_{\frac{3R+\varepsilon}{2}}(u) \cup B_{\frac{3R+\varepsilon}{2}}(v).$$

As $\|u-v\|> \frac{3R+\varepsilon}{2}+\frac{3R+\varepsilon}{2}+3\varepsilon$ we further conclude that $[x]$ is contained in only one of $B_{\frac{3R+\varepsilon}{2}}(u)$ or $B_{\frac{3R+\varepsilon}{2}}(v)$. □

In light of Proposition 1.3.4 we modify our partition of P, into $\widetilde{P_0}$ and $\widetilde{P_1}$, see Definition 1.3.5 and Algorithm 2. We effectively want to move any points in P_1 that are not contained in a spanning connected component into P_0.

Definition 1.3.5 ($\widetilde{P_0}$ and $\widetilde{P_1}$)**.** *Let P be an ε-sample of an embedded graph $|G|$ satisfying Assumptions 1, and consider the sets P_0 and P_1 from Definition 1.2.5. Let Q_0 be the connected components of $\mathfrak{G}_{\frac{3R}{2}+2\varepsilon}(P_0)$, and Q_1 the connected components of $\mathfrak{G}_{3\varepsilon}(P_1)$, and define $f : Q_1 \to \{0,1\}$ by $f([q]) = 0$ when there is only a single connected component $[p] \in Q_0$ such that $d([p],[q]) < 3\varepsilon$, and $f([q]) = 1$ otherwise.*

We define $\widetilde{P_0} := P_0 \cup \left(\bigcup_{f([x])=0}[x]\right)$ and $\widetilde{P_1} := \left(\bigcup_{f([x])=1}[x]\right)$.

Lemma 1.3.6. *Let $[x] \in Q_1$. Then $f([x]) = 1$ if and only is $[x]$ spans an edge, and $f([x]) = 0$ if and only if $[x] \subset B_{\frac{3R+\varepsilon}{2}}(v)$ for some vertex v.*

Proof. If $[x]$ spans an edge $\overline{uv}$ then by Proposition 1.3.2 we know that $[x]$ contains samples in both $B_{\frac{3R+5\varepsilon}{2}}(u)$ and $B_{\frac{3R+5\varepsilon}{2}}(v)$. Let $x_u \in [x]$ be the sample closest to u. Note that $\|x_u - u\|\leq \frac{3R+5\varepsilon}{2}$. There must be some sample $p_u \in P$ with $\|p-u\|\in< \|u-x_u\|$ and $\|p-x_u\|< 3\varepsilon$. Now $p \in P_0$ as otherwise it contradicts x_u being the closest sample to u inside $[x]$. By Lemma 1.3.1, $[p_u] \in Q_0$ is contained in $B_{\frac{3R+\varepsilon}{2}}(u)$.

Similarly we can show that there some $x_v \in [x]$ and $p_v \in P_0$ with $\|x_v - p_v\|\leq 3\varepsilon$ and $[p_v] \in Q_0$ contained in $B_{\frac{3R+\varepsilon}{2}}(v)$. By Lemma 1.3.1 $[p_u]$ and $[p_v]$ are distinct and hence $f([x]) = 1$.

If $[x]$ does not span any edge then by Proposition 1.3.2 we know there is a vertex v such that $[x] \subset B_{\frac{3R+\varepsilon}{2}}(v)$. We then can appeal to Lemma 1.3.1 to say that there is only one connected component in Q_0 within 3ε of $[x]$. □

Let $\widetilde{Q_0}$ denote the connected components of $\mathfrak{G}_{\frac{3R}{2}+2\varepsilon}(\widetilde{P_0})$ and let $\widetilde{Q_1}$ denote the connected components of $\mathfrak{G}_{3\varepsilon}(\widetilde{P_1})$. We will see that characterisation of the elements of $\widetilde{Q_0}$ is the same as that of Q_0. The elements of $\widetilde{Q_1}$ are exactly those connected components that span some edge.

Theorem 1.3.7. *For each vertex v there exists a unique connected component $[x] \in \mathfrak{G}_{\frac{3R}{2}+2\varepsilon}(\widetilde{P_0})$ such that $[x] \subset B_{\frac{3R}{2}+2\varepsilon}(v)$. Every connected component of $\mathfrak{G}_{\frac{3R}{2}+2\varepsilon}(\widetilde{P_0})$ is of this form.*

For each edge $\overline{uv}$ there exists a unique connected component $[x] \in \mathfrak{G}_{3\varepsilon}(\widetilde{P_1})$ such that $[x]$ spans $\overline{uv}$. Furthermore every connected component of $\mathfrak{G}_{\frac{3R}{2}+2\varepsilon}(\widetilde{P_1})$ is of this form.

Proof. From Proposition 1.2.11 and Lemma 1.3.6 we know that $\widetilde{P_0} \subset \bigcup_v B_{\frac{3R+\varepsilon}{2}}(v)$. We can then effectively repeat the proof of Lemma 1.3.1 to show the analogous result for $\widetilde{P_0}$.

To see the bijection between the vertices of G and $\widetilde{Q_0}$ observe that every sample within 4ε of some vertex is in $P_0 \subset \widetilde{P_0}$ and hence every vertex corresponds to some connected component, and observe that by Lemma 1.3.6 all points in $\widetilde{P_0}$ lie within $\frac{3R+\varepsilon}{2}$ of some vertex.

The characterisation for connected components of $\mathfrak{G}_{3\varepsilon}(\widetilde{P_1})$ follows directly from Proposition 1.3.2 and Lemma 1.3.6. □

Define the map $F_0 : \widetilde{Q_0} \to V$ by $F_0([x]) = \text{argmin}_{v\in V}\{d([x], v)\}$ and $F_1 : \widetilde{Q_1} \to E$ by $F_1([x]) = \text{argmin}_{\overline{uv}\in E}\{d([x], \text{midpt}(\overline{uv})\}$.

That F_0 and F_1 are well-defined bijections follows directly from Theorem 1.3.7. From Proposition 1.3.2 we further can say that if $[q] \in \widetilde{Q_1}$ and $[x] \in \widetilde{Q_0}$ then the single linkage distance between $[q]$ and $[x]$ is less than 3ε if and only if $F_0([x]) \in \partial_G(F_1([q]))$.

1.4. Vertex prediction

Thus far, the focus has been on finding the abstract structure of an embedded graph $|G|$. We now aim to form a numerical scheme to estimate the vertex locations of $|G|\subset \mathbb{R}^n$. In [4], a non-linear least-squares method was proposed and used for embedded graph reconstruction. Empirical observation of this method showed vertex predictions were often not within ε of the *true* embedded graph. A point of difficulty here was that data that should belong to a one-dimensional strata piece was often assigned to a zero-dimensional strata when nearby a vertex location. We utilise an Expectation-Maximisation (EM) algorithm, which updates both the predicted vertex locations, and their strata assignments to correct this issue.

Algorithm 1: $\Delta_{R,\varepsilon}(p)$

```
Data: An $\varepsilon$-dense sample $P$ of an embedded graph $|G|$, a point
      $p \in P$.
Result: 0 if $p$ has local structure of a vertex, 1 if $p$ has local
        structure of an edge.
begin
    $\mathcal{G}_p \longleftarrow \{q \in P \mid \|p - q\| \leq R + \varepsilon\}$;
    connect $q, q' \in \mathcal{G}_p$ if $\|q - q'\| \leq 3\varepsilon$;
    if $\mathcal{G}_p$ is disconnected then
        return 1
    else
        remove $q \in \mathcal{G}_p$ if $\|p - q\| \leq R - \varepsilon$;
        if number of connected components in $\mathcal{G}_p$ is not $2$ then
            return 0
        else
            find the midpoints $q_1, q_2$ of the connected components
              $c_1$ and $c_2$;
            if $\langle q_1 - p, q_2 - p\rangle > -R^2 + 2R\varepsilon - 7\varepsilon^2$ then
                return 0
            else
                return 1
```

To do this, we design a likelihood function with latent variables for strata assignment so that we may reconstruct a probability measure over the embedded graph from which our data is sampled. Ideally, we would reconstruct a measure ν whose support is the embedded graph. Recorded data has errors and makes it computationally infeasible to reconstruct ν directly. Instead, we will formulate an approximating measure ν_δ which satisfies:

1. ν_δ is equivalent to Lebesgue measure,
2. $\text{supp}(\lim_{\delta \to 0} \nu_\delta) = |G|$,

where the limit is meant in the weak sense. The first assumption gives robustness to measurement errors, and the second ensures that in ideal circumstances, we form a measure that is supported on the embedded graph. There are many measures which obey these conditions, we choose a Gaussian convolution model for each strata piece and combine all the strata pieces together through a categorical mixture model.

```
Algorithm 2: Abstract Structure
  Data: Partition of $P$ into $P_0$ and $P_1$.
  Result: Partitions $\widetilde{P_0}, \widetilde{P_1}$, abstract graph $G = (E, V)$.
  begin
    $E \longleftarrow \emptyset$;
    $V \longleftarrow \emptyset$;
    $\widetilde{P_0} \longleftarrow P_0$;
    $\widetilde{P_1} \longleftarrow P_1$;
    for connected components $[p] \in \mathfrak{G}_{\frac{3R}{2}+2\varepsilon}(P_0)$ do
      add $[p]$ to $V$
    for connected components $[q] \in \mathfrak{G}_{3\varepsilon}(P_1)$ do
      $B_q \longleftarrow \emptyset$;
      for $[p] \in V$ do
        if $\min_{p' \in [p], q' \in [q]} \|p' - q'\| \leq 3\varepsilon$ then
          add $[p]$ to $B_q$
    if size$(B_q) = 1$ then
      add all $q' \in [q]$ to $\widetilde{P_0}$ and remove them from $\widetilde{P_1}$
    else
      add $B_q$ to $E$
  return $\widetilde{P_0}, \widetilde{P_1}, V, E$
```

1.4.1. Embedded graph model. Let $(\Omega, \mathcal{F}, \mu)$ be a probability space, that is Ω is a set, $\mathcal{F}$ is a σ-algebra of sets from Ω, and $\mu : \mathcal{F} \mapsto [0, 1]$ is a normalised measure.

Definition 1.4.1. Given a probability space $(\Omega, \mathcal{F}, \mu)$ and a field with a σ-algebra $\mathcal{B}$, a measurable function $f : (\Omega, \mathcal{F}, \mu) \mapsto (\mathbb{F}, \mathcal{B})$ is a *random variable*. A vector valued random element is a vector valued measurable function $\tilde{f} : (\Omega, \mathcal{F}, \mu) \mapsto (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ given through $\tilde{f} = (f_1, \ldots, f_n)$ where each of the f_i are random variables.

The *expectation* of a random variable is the integral, $\mathbb{E}(f) := \int_\Omega f d\mu$. Given a sub-$\sigma$-algebra $C \subset \mathcal{F}$, the *conditional expectation* of a random variable f, $\mathbb{E}(f|C) \in L^2(\Omega, \mathcal{F}, \mu)$, is the unique function that satisfies

$$\int_B \mathbb{E}(f|C) d\mu = \int_B f d\mu,$$

Algorithm 3: Expectation Maximisation for Vertex Location Prediction

```
Data: $|P|$ data points in $n$ dimensions, $N_0 + N_1 = N$ many strata pieces.
Result: Predicted embedded graph vertex locations.
Input: Abstract graph structure.
begin
    Initialise vertex locations $V$ ;
    Initialise $|P| \times N$ strata assignment matrix $A$ ;
    for $s_i$ in strata pieces $S = V \cup E$, $x_j$ in data points do
        if $x_j \in s_i$ then
            $A_{i,j} \longleftarrow 1$
        else
            $A_{i,j} \longleftarrow 0$
    assign an error threshold $\sigma \in \mathbb{R}_+$;
    Initialise $\pi_i = \frac{\sum_i A_{i,j}}{\sum_{i,j} A_{i,j}}$;
    for iterations in EM-iterations do
        for $s_i$ in strata pieces $S = V \cup E$, $x_j$ in data points do
            assign $A_{i,j} = \mathbb{E}(1_{Z_j=1} | X_j = x_j)$ through (1.10) ;
        assign $\pi_i = \frac{\sum_i A_{i,j}}{\sum_{i,j} A_{i,j}}$;
        assign $V = \arg\min_V V \to C(V, \Pi; \sigma)$ (1.9) through a hill
          climbing optimiser such as gradient-descent;
```

for all $B \in C$. The expectation and conditional expectation of a vector valued random element $\tilde{f} = (f_1, \ldots, f_n)$ is defined component-wise through each of the random variables f_i, that is

$$\mathbb{E}(\tilde{f}|C) := (\mathbb{E}(f_1|C), \ldots, \mathbb{E}(f_n|C)) \tag{1.6}$$

for all $C \in \mathcal{B}(\mathbb{R}^n)$.

Above, we have adopted the standard notation $\mathcal{B}(\mathbb{R}^n)$ for the Borel-σ-algebra generated by the open sets in the standard topology on $\mathbb{R}^n$. Let $X_j : (\Omega, \mathcal{F}, \mu) \mapsto (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ be vector valued random elements and $Z_j : (\Omega, \mathcal{F}, \mu) \mapsto ([N], 2^{[N]})$ be random variables for $j \in \{1, \ldots, |P|\}$, where $[N] := \{1, \ldots, N\}$, n is the dimension of the space to which the graph is embedded, $|P|$ is the amount of recorded data points, and N the number of strata in $|G|$. Let $N_0, N_1 \in \mathbb{N}_0$ be N_0 and N_1 are the number of zero and one dimensional strata respectively, and so $N = N_0 + N_1$.

Enumerate the set of vertex locations as $V := \{v_i\}_{i=1}^{N_0}$. For each $i \in \{N_0 + 1, \dots, N\}$ assign the pairing $v_{i_1}, v_{i_2} \in \{v_i\}_{i=1}^{N_0}$ to be the vertices that form the boundary i^{th} strata piece. Assume that for each j the Z_j are independent and identically distributed, that each X_j is independent of X_i and Z_i for $i \neq j$.

We place the following constraints on the random variables:

1. $Z_j \sim$ Categorical(Π) with parameters $\Pi := (\pi_1, \dots, \pi_N)$,
2. $\mathbb{E}(X_j|Z_j = i) \sim$ Normal(v_j, σ_j) for $j \in \{1, \dots, N_0\}$,
3. $\mathbb{E}(X_j|Z_j = i) = t_j v_{i_1} + (1 - t_j)v_{i_2} + \varepsilon$ where $t_j \sim$ Uniform$([0, 1])$ and
   $\varepsilon_j \sim$ Normal$(0, \sigma_i)$ for $i \in \{N_0 + 1, \dots, N\}$.

The categorical random variables Z_j represent which stratum a random element X_j belongs to. The categorical distribution is defined on N many categories, with the i^{th} category having a probability of π_i of being observed. In our case, each π_i represents approximately how many data points belong to the i^{th} stratum.

The distribution of $\mathbb{E}(X_j|Z_j = i \in \{N_0 + 1, \dots, N\}) = tv_{j_1} + (1 - t)v_{j_2} + \varepsilon$ is

$$\rho_{v_{i_1}, v_{i_2}}(x; \sigma_i) := \frac{1}{(2\pi\sigma_i^2)^{n/2}} \int_0^1 e^{-\|x-(tv_{i_1}+(1-t)v_{i_2})\|_2^2/2\sigma_i^2} dt, \tag{1.7}$$

where $\rho(\,\cdot\,; 0, \sigma_i)$ is a normal density in n dimensions with zero mean and variance σ_i^2. This can be obtained through noting that if $\nu_{v_{i_1}, v_{i_2}}$ is uniform measure on $\mathcal{L}_{v_{i_1}, v_{i_2}} := \{y \mid y = tv_{i_1} + (1 - t)v_{i_2},\ t \in [0, 1]\}$, then the measure

$$\nu_{\sigma_i, v_{i_1}, v_{i_2}} = \rho_{v_{i_1}, v_{i_2}}(x; \sigma_i)dx$$

is given through

$$\begin{aligned}\nu_{\sigma_i, v_{i_1}, v_{i_2}} &= \rho(x; 0, \sigma_i)dx * \nu_{v_{i_1}, v_{i_2}} \\ &= \frac{1}{(2\pi\sigma_i^2)^{n/2}} \int_0^1 e^{-\|x-(tv_{i_1}+(1-t)v_{i_2})\|_2^2/2\sigma_i^2} dt dx,\end{aligned}$$

where $*$ represents the convolution operation over measures. Through this convolution construction, we have the following proposition.

Proposition 1.4.2. *Let $\rho_{v_{i_1},v_{i_2}}$ and $\nu_{v_{i_1},v_{i_2}}$ be as given above, then:*

1. *$\rho_{v_{i_1},v_{i_2}} \in C^\infty(\mathbb{R}^n)$,*
2. *$\rho_{v_{i_1},v_{i_2}} dx$ is equivalent to Lebesgue measure,*
3. *and $\nu_{\sigma_i,v_{i_1},v_{i_2}} \xrightarrow{\sigma_i \to 0} \nu_{v_{i_1},v_{i_2}}$ weakly.*

The first two claims follow from Equation 1.7. The third is a result from mollifier approximation theory, see [21] for details.

Corollary 1.4.3. *Define $\sigma := \max_i \sigma_i$ and let $\nu_\sigma := \mu(X_j^{-1})$ be the push-forward measure of μ through X_j, then*

1. *$\nu_\sigma \sim dx$,*
2. *supp$(\lim_{\sigma \to 0} \nu_\sigma) = |G|$,*
3. *$\lim_{\sigma \to 0} \nu_\sigma(|G|) = 1$,*

where $|G|$ is the embedded graph in $\mathbb{R}^n$.

Proof. Write ν_σ through

$$\begin{aligned}\nu_\sigma &= \sum_{i=1}^{N_0} \pi_i \rho(x; v_i, \sigma_i) dx + \sum_{i=N_0+1}^{N} \pi_i \rho_{v_{j_1},v_{j_2}}(x;\sigma_i) dx \\ &= \sum_{i=1}^{N_0} \pi_i (\delta_{v_i} * \rho(x; 0, \sigma_i) dx) + \sum_{i=N_0+1}^{N} \pi_i (\nu_{v_{i_1},v_{i_2}} * \rho(x; 0, \sigma_i) dx)\end{aligned}$$

where δ_{v_i} is the normalised measure: $\delta_{v_i}(U) = 1$ if $v_i \in U$ and zero otherwise. Let $|G|$ be the embedded graph and define $|G|_r := \{x \mid x \in B_r(y),\ y \in |G|\}$, then

$$\begin{aligned}\nu_\sigma(\mathbb{R}^n \setminus |G|_r) \leq \int_{\mathbb{R}^n \setminus |G|_r} &\left(\sum_{i=1}^{N_0} \pi_i \delta_{v_i} + \sum_{i=N_0+1}^{N} \pi_i \nu_{v_{i_1},v_{i_2}} \right) * \rho(x; 0, \sigma) dx \\ &\xrightarrow{\sigma \to 0} 0 \quad \text{for all } r > 0.\end{aligned}$$

□

Corollary 1.4.3 shows the push-forward measure ν has our desired properties for modelling an embedded graph $|G|$.

1.4.2. Parameter re-estimation. We now form an Expectation Maximisation (EM) algorithm to find Maximum Likelihood Estimates (MLEs) for the embedded graph's vertex locations. Let $\widetilde{\mathbb{P}}(\Omega)$ be the space of probability measures over Ω. We are interested in reconstructing the measure μ given evaluations of X_j and Z_j for every $j \in \{1, \dots, n\}$. This forms the following likelihood optimisation problem:

$$\mu^* := \text{argsup}_{\eta \in \widetilde{\mathbb{P}}(\Omega)} \eta \left(\bigcap_{j \in \{1, \dots, |P|\}} X_j^{-1}(B_h(x_j)) \cap Z_j^{-1}(i) \right)$$

for some small $h > 0$. For a single recorded datum:

$$\eta(X_j^{-1}(B_h(x_j)) \cap Z_j^{-1}(i)) = \mathbb{P}(X_j \in B_h(x_j) \mid Z_j = i)\mathbb{P}(Z_j = i)$$
$$= \prod_{i=1}^{N_0} \left(\pi_i \int_{B_h(x_j)} \rho(x; v_i, \sigma_i) dx \right)^{1_{Z_j=i}} \prod_{i=N_0+1}^{N} \left(\pi_i \int_{B_h(x_j)} \rho_{v_{j_1}, v_{j_2}}(x; \sigma_i) dx \right)^{1_{Z_j=i}} .$$

Intersecting over all such data points, taking a logarithm, and evaluating the limit as $h \to 0$ for the argument supremum yields the equivalent optimisation:

$$\text{argsup}_{\pi_i \in [0,1],\ v_i \in \mathbb{R}^n} \sum_{j=1}^{|P|} \Big(\sum_{i=1}^{N_0} 1_{Z_j=i} (\log(\rho(x_j; v_i, \sigma_i)) + \log(\pi_i)) + \sum_{i=N_0+1}^{N} 1_{Z_j=i} (\log(\rho_{v_{i_1}, v_{i_2}}(x_j; \sigma_i)) + \log(\pi_i)) \Big). \tag{1.8}$$

We cannot observe accurately Z_j for a recorded datum, although the work in estimating the abstract graph structure gives an initial estimate for this value. To dynamically update the prediction of this value, we will utilise an EM-algorithm. Projection to the sub-σ-algebra $\sigma(X_1, \dots, X_n)$ and making the assumption $Z_j \perp X_{\widetilde{j}}$ for $\widetilde{j} \neq j$ gives the following log-likelihood function, which we aim to maximise:

$$\begin{aligned}
&\mathcal{L}(V, \Pi; \sigma) := \\
&\quad \frac{1}{|P|}\sum_{j=1}^{|P|}\Big(\sum_{i=1}^{N_0}\mathbb{E}(1_{Z_j=i}|X_j \in B_{h'}(x_j))(\log(\rho(x_j; v_i, \sigma_i)) + \log(\pi_i)) + \\
&\qquad \sum_{i=N_0+1}^{N}\mathbb{E}(1_{Z_j=i}|X_j \in B_{h'}(x_j))(\log(\rho_{v_{i_1},v_{i_2}}(x_j; \sigma_i)) + \log(\pi_i))\Big),
\end{aligned} \tag{1.9}$$

where we currently view $\mathcal{L}$ as a function of the vertex locations V and assignment weights Π, with σ being a fixed value. Let the densities for each $k \in \{1, \dots, N\}$ strata be enumerated as $\{\rho_k\}_{k=1}^{N}$. The individual terms of the cost function are

$$\begin{aligned}
\lim_{h' \to 0}\mathbb{E}(1_{Z_j=i}|X_j \in B_{h'}(x_j)) &= \frac{\pi_i \rho_i(x_j)}{\sum_{k=1}^{N}\pi_k \rho_k(x_j)} \\
\log(\rho(x; v_i, \sigma_i)) &= -\frac{d}{2}\log(2\pi\sigma_i) - \|x - v_i\|^2/2\sigma_i.
\end{aligned} \tag{1.10}$$

$$\begin{aligned}
\log(\rho_{v_{i_1},v_{i_2}}(x; \sigma_i)) = \log\Bigg(&\text{erf}\left(\frac{\langle v_{i_1} - v_{i_2}, v_{i_1} + v_{i_2} - 2x\rangle + \|v_{i_1} - v_{i_2}\|_2^2}{2\sqrt{2}\|v_{i_1} - v_{i_2}\|_2\sigma_i}\right) \\
&- \text{erf}\left(\frac{\langle v_{i_1} - v_{i_2}, v_{i_1} + v_{i_2} - 2x\rangle - \|v_{i_1} - v_{i_2}\|_2^2}{2\sqrt{2}\|v_{i_1} - v_{i_2}\|_2\sigma_i}\right)\Bigg) \\
&+ \frac{\langle v_{i_1} - v_{i_2}, v_{i_1} + v_{i_2} - 2x\rangle^2 - 4\|v_{i_1} - v_{i_2}\|_2^2\|(v_{i_1} + v_{i_2})/2 - x\|_2^2}{8\|v_{i_1} - v_{i_2}\|_2^2\sigma_i^2} \\
&- \log(\|v_{i_1} - v_{i_2}\|_2) + \log\left(2^{\frac{1}{2}(-d-1)}\pi^{\frac{1}{2}-\frac{d}{2}}\sigma_i^{1-d}\right).
\end{aligned}$$

Above, erf : $\mathbb{R} \mapsto \mathbb{R}$ is the standard error function given through

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x \exp(-t^2)dt.$$

In Skyler, the analytic gradients of the log-likelhood function $\mathcal{L}$ are given. Gradient clipping is used to bound our computations within machine accuracy for when σ_i or the evaluation of $x \mapsto \rho_{v_{i_1},v_{i_2}}(x; \sigma_i)$ is close to machine precision. Our log-likelihood function is often not concave, for instance the function $(v_{i_1}, v_{i_2}) \mapsto \rho_{v_{i_1},v_{i_2}}(x; \sigma_i)$ obeys $\rho_{v_{i_1},v_{i_2}}(x; \sigma_i) = \rho_{v_{i_2},v_{i_1}}(x; \sigma_i)$.

It is necessary to have a good initialisation for the embedded graph modelling to find an acceptable local optimum value for vertex prediction. In our computations, we have found that the initial vertex modelling given by the abstract graph structure yields vertex predictions with an error less than the noise of the data, correcting the issue observed in [4]. We can complete Algorithm 3 by noting that if $A_{i,j} := \lim_{h' \to 0} \mathbb{E}(1_{Z_j = i} | X_j \in B_{h'}(x_j))$, then the function $\Pi \to \mathcal{L}(V, \Pi; \sigma)$ is concave and has a unique maximum value at $\pi_i^* = \frac{\sum_j A_{i,j}}{\sum_{i,j} A_{i,j}}$. It can be seen that our model is a higher-dimension version of Gaussian clustering as Algorithm 3 degenerates to this when $N = N_0$.

Fixing a noise tolerance σ and solving the optimisation in Equation 1.8 by minimising the function $(V, \Pi) \to \mathcal{L}(V, \sigma, \Pi)$ through an EM-algorithm [12] gives Algorithm 3.

1.4.3. Numerical simulations. The conditions in Assumption 1 are not the sharpest bounds, and other ratios of R and ε can also detect the correct graph structure. We present the results of a few different ratios, for the same 0.1-sample P (Figure 1.9B) of the embedded graph $(G, \phi_G) \subset \mathbb{R}^3$ (Figure 1.9A). There are 705 samples in P, and G has 5 vertices embedded as 1: $(0,0,0)$, 2: $(4.6, 6.24, 0)$ 3: $(4.86, 0.51, 3.47)$, 4: $(-1.32, 6.29, 4)$, and 5: $(-4.23, -3.48, -3)$, and edges $E = \{(1,5), (1,3), (1,4), (2,4), (2,3)\}$.

Table 1 shows the results with varying choices of ratio $\frac{R}{\varepsilon}$. Comparing the log-likelihood of the models obtained using $\frac{R}{\varepsilon} = 8\,(-2.3712314714356437)$ and $\frac{R}{\varepsilon} = 12\,(-2.783827546761547)$, we see that while we have shown that $R \geq 12\varepsilon$ is sufficient to prove correctness of the algorithm, smaller ratios can also identify an isomorphic graph structure, and result in a higher log-likelihood model. In practice, this suggests that we can improve the process by first using $R \geq 12\varepsilon$ to obtain the correct structure, and then decreasing the ratio to model the graph, stopping when we still obtain the correct graph structure and maximise the log-likelihood.

1.5. Future directions

The algorithm presented in this chapter focuses on recovering and modelling an embedded graph (G, ϕ_G) given an ε-sample P. Stratified spaces, however, are not restricted to consisting of 0- and 1-dimensional pieces, nor

Ratio	Correct	Log Likelihood	v_1	v_2	v_3	v_4	v_5
R/ε	structure	(Equation 1.9)	$\begin{pmatrix}0\\0\\0\end{pmatrix}$	$\begin{pmatrix}4.6\\6.24\\0\end{pmatrix}$	$\begin{pmatrix}4.86\\0.51\\3.47\end{pmatrix}$	$\begin{pmatrix}-1.32\\6.29\\4\end{pmatrix}$	$\begin{pmatrix}-4.23\\-3.48\\-3\end{pmatrix}$
4	No	-	-	-	-	-	-
6	Yes	-33.183	$\begin{pmatrix}0.00\\0.03\\0.01\end{pmatrix}$	$\begin{pmatrix}4.59\\6.23\\-0.02\end{pmatrix}$	$\begin{pmatrix}4.56\\0.56\\3.43\end{pmatrix}$	$\begin{pmatrix}-1.30\\6.26\\3.96\end{pmatrix}$	$\begin{pmatrix}-4.24\\-3.46\\-3.02\end{pmatrix}$
8	Yes	-32.97	$\begin{pmatrix}0.00\\0.02\\0.01\end{pmatrix}$	$\begin{pmatrix}4.59\\6.23\\-.02\end{pmatrix}$	$\begin{pmatrix}4.86\\0.56\\3.42\end{pmatrix}$	$\begin{pmatrix}-1.29\\6.26\\3.96\end{pmatrix}$	$\begin{pmatrix}-4.22\\-3.45\\-3.01\end{pmatrix}$
10	Yes	-33.33	$\begin{pmatrix}0.01\\0.03\\0.01\end{pmatrix}$	$\begin{pmatrix}4.59\\6.22\\-0.02\end{pmatrix}$	$\begin{pmatrix}4.86\\0.56\\3.42\end{pmatrix}$	$\begin{pmatrix}-1.26\\6.24\\3.95\end{pmatrix}$	$\begin{pmatrix}-4.19\\3.42\\-2.99\end{pmatrix}$
12	Yes	-33.84	$\begin{pmatrix}0.01\\0.03\\0.01\end{pmatrix}$	$\begin{pmatrix}4.59\\6.23\\-0.02\end{pmatrix}$	$\begin{pmatrix}4.86\\0.56\\3.42\end{pmatrix}$	$\begin{pmatrix}-1.26\\6.24\\3.95\end{pmatrix}$	$\begin{pmatrix}-4.14\\-3.38\\-2.96\end{pmatrix}$
14	Yes	-36.61	$\begin{pmatrix}0.01\\0.03\\0.01\end{pmatrix}$	$\begin{pmatrix}4.59\\6.26\\-0.03\end{pmatrix}$	$\begin{pmatrix}4.86\\0.56\\3.43\end{pmatrix}$	$\begin{pmatrix}-1.26\\6.23\\3.95\end{pmatrix}$	$\begin{pmatrix}-4.00\\-3.27\\-2.56\end{pmatrix}$
16	Yes	-45.30	$\begin{pmatrix}0.02\\0.03\\0.01\end{pmatrix}$	$\begin{pmatrix}4.58\\6.27\\-0.05\end{pmatrix}$	$\begin{pmatrix}4.56\\0.56\\3.43\end{pmatrix}$	$\begin{pmatrix}-0.70\\3.70\\2.33\end{pmatrix}$	$\begin{pmatrix}-3.96\\-3.22\\-2.81\end{pmatrix}$

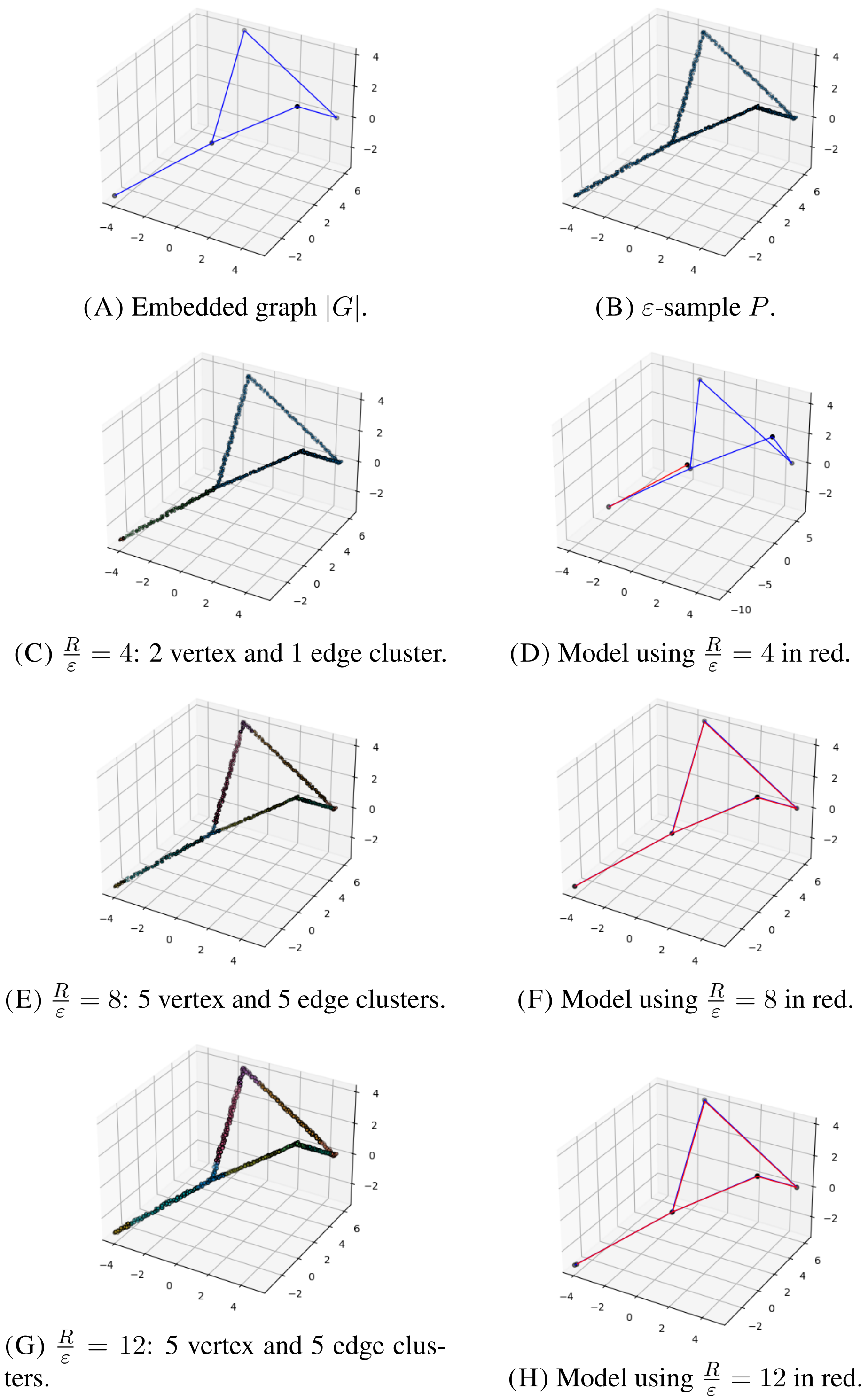
TABLE 1. Summary of the output of the algorithm for various ratios $\frac{R}{\varepsilon}$. Recall we wish to maximise Equation 1.9. The last 5 columns are the vertex locations obtained.

are they restricted to being simplicial complexes. We can consider embeddings of CW complexes, where a stratum is embedded as a semi-algebraic set.

While the algorithm in this chapter does not naively extend to higher simplicial complexes or CW complexes, it provides a foundation on which other algorithms can be based, and hence moves towards learning general stratified spaces. The algorithm can be adapted to other cases and assumptions. For example, it can be adapted to learn the abstract structure of a graph with non-linear edges and no degree 2 vertices. In particular, to recover embedded CW complexes, we need to remove the assumption that strata are embedded as convex hulls (linearity).

Focusing on increasing the dimension of the cells in the simplicial complex, the next step is to allow 2-simplicies and partition an ε-sample P into three parts P_0, P_1, and P_2. One approach is a peeling argument: first we determine the points in P_2, and then apply the current algorithm to $P \setminus P_2$ to obtain P_1 and P_0. Complications with this include ensuring that points are not over-assigned to $P \setminus P_2$, as this can result in $P \setminus P_2$ not being suitable as input for the current algorithm. To appropriately partition P, we hope to exploit the relationship between (R, ε)-local structure and local homology. For graphs, we saw that the dimension 1 local homology at a point x contains topological information, which corresponds to the number of points in the intersection of the $|G|$ with a ball of small radius r around x, and if there are 2 points, their relative geometry providing more information. By generalising the (R, ε)-local structure appropriately, we hope to see a correspondence with the information contained in higher homology groups and augment this with other geometrical information.

To remove the linearity assumption, we need to address a long standing problem in computational algebraic geometry: learning algebraic varieties from noisy samples. In [6], Breiding et al. develop an algorithm which is robust to machine error but not sampling noise. The algorithm has also been found to fail when given large data sets sampled from simple varieties. These issues need to be overcome before we can remove the linearity assumption.



(A) Embedded graph $|G|$.

(B) ε-sample P.

(C) $\frac{R}{\varepsilon} = 4$: 2 vertex and 1 edge cluster.

(D) Model using $\frac{R}{\varepsilon} = 4$ in red.

(E) $\frac{R}{\varepsilon} = 8$: 5 vertex and 5 edge clusters.

(F) Model using $\frac{R}{\varepsilon} = 8$ in red.

(G) $\frac{R}{\varepsilon} = 12$: 5 vertex and 5 edge clusters.

(H) Model using $\frac{R}{\varepsilon} = 12$ in red.

FIGURE 1.9

CHAPTER 2

Learning 2-complexes

> Perhaps home is not a place but simply an irrevocable condition.
>
> James Baldwin, *Giovanni's Room*

Chapter 1 presents an algorithm for learning an abstract graph G and modelling an embedding $|G|\subset \mathbb{R}^n$ from an ε-sample $P \subset \mathbb{R}^n$ of $|X|$. A natural question is how to extend Algorithms 1 and 2 to learn embeddings $|X|\subset \mathbb{R}^n$ of abstract complexes X from ε-samples $P \subset \mathbb{R}^n$ of $|X|$. This requires identifying the total dimension of X, and the *locally* top dimensional cells. In this chapter, we restrict to 2-complexes.

This chapter begins with Section 2.1, containing definitions of the main objects and tools we use throughout the chapter. After this, Section 2.2 consists of geometric lemmas used in Section 2.3, which considers the local geometry and topology we use to partition the sample P. Finally, Section 2.4 presents algorithms for recovering the abstract structure. Section 2.4 contains a sequence of lemmas (Lemmas 2.4.10 to 2.4.25), which cover cases used in Theorem 2.4.26, also known as the 'Big Theorem' of this chapter.

2.1. Definitions and Notations

We begin with some definitions and notations we use throughout this article. We begin with the following definition of complex, following Definition 2.4 [8].

Definition 2.1.1 (Abstract Complex, Definition 2.4 [8])**.** *An* abstract simplicial complex X *consists of a pair* $(V(X), \Sigma(X))$*, with* $V(X)$ *a finite set, and* $\Sigma(X)$ *a subset of the power set of* $V(X)$*, such for all* $\sigma \in \Sigma(X)$ *and* $\emptyset \neq \tau \subseteq \sigma$*, we have* $\tau \in \Sigma(X)$*. We call* $V(X)$ *the vertices, and* $\Sigma(X)$ *the* simplices *of* X.

For ease of notation and to avoid confusion later in this paper, we will use the following specialised definition for abstract simplicial complexes with top dimension 2.

Definition 2.1.2 (Abstract 2-Complex)**.** *An* abstract 2-complex X *consists of*

1. *a set* $V = V(X)$ *of vertices,*
2. *a set* $E = \{\, \sigma \in \Sigma(X) \,|\, \sigma$ *contains 2 unique elements*$\}$ *of edges,*
3. *a set* $T = \{\, \sigma \in \Sigma(X) \,|\, \sigma$ *contains 3 unique elements*$\}$ *of triangles,*

and an incidence operator $\mathcal{I}$*, which acts as follows: for any pair of cells* $\sigma, \tau \in X$

$$\mathcal{I}\,(\sigma, \tau) = \begin{cases} 1 & \textit{if } \sigma \subsetneq \tau \\ 0 & \textit{otherwise} \end{cases}$$

We restrict ourselves to linear embeddings of 2-complexes X in $\mathbb{R}^n$ for some $n \geq 3$.

Definition 2.1.3 (Linear embedding of 2-complex)**.** *Fix* $n \geq 3$*, then a linear embedding of a* 2*-complex* X *in* $\mathbb{R}^n$*,* (X, Θ)*, consists of an abstract* 2*-complex* X *and a map*

$$\Theta : X \to \mathbb{R}^n$$

such that

1. *on vertices* $v \in V$*,* Θ *is injective,*
2. *on edges* $\{u, v\} \in E$*,* Θ *is defined by linear interpolation on* $\Theta(u)$ *and* $\Theta(v)$*:* $\Theta(\{\, u, v \,\}) = \overline{uv}$ *is the line segment between* $\Theta(u)$ *and* $\Theta(v)$*,*
3. *on triangles* $\{\, u, v, w \,\} \in E$*,* Θ *is defined by linear interpolation on* $\Theta(u)$*,* $\Theta(v)$ *and* $\Theta(w)$*:* $\Theta(\{\, u, v, w \,\}) = \triangle uvw$ *is the triangle with vertices* $\Theta(u)$*,* $\Theta(v)$ *and* $\Theta(w)$*, and* $\Theta(u)$*,* $\Theta(v)$*,* $\Theta(w)$ *are no co-linear,*
4. *for any two cells* σ, τ *of* X*, we have* $\Theta(\sigma) \cap \Theta(\tau) = \Theta(\sigma \cap \tau)$*.*

We restrict our attention to embedded 2*-complexes* $|X|_\Theta$ *such that*

5. *if a vertex* v *is in the boundary of precisely two edges* $\{v, u_1\}$ *and* $\{v, u_2\}$*, then* $\angle u_1 v u_2 \neq \pi$*,*

6. *if an edge $\{v_0, v_1\}$ is in the boundary of precisely two triangles $\{v_0, v_1, u_1\}$ and $\{v_0, v_1, u_2\}$, then v_0, v_1, u_1, u_2 are not co-planar.*

We denote the image of Θ in $\mathbb{R}^n$ by $|X|_\Theta$.

We often talk about the *boundary* of a cell.

Definition 2.1.4 (Cell boundary)**.** *Let X be an abstract 2-complex, and take a cell $\sigma \in X$. Then the* boundary *of τ, $\partial\tau$, consists of the cells $\sigma \in X$ such that $\mathcal{I}(\sigma, \tau) = 1$.*

An important property of a cell $\sigma \in X$, is whether it is *locally maximal* or not.

Definition 2.1.5 (Locally maximal cell)**.** *Let σ be a cell in a 2-complex. We say σ is* locally maximal *if there is no cell $\tau \in X, \tau \neq \sigma$ with $\sigma \subset \tau$. That is, there is no cell τ with σ in the boundary of τ.*

Remark 2.1.6. Consider two cells σ, τ in a complex X, we say σ *is a face of* τ if σ is in the boundary of τ, and we say σ *is a co-face of* τ if τ is in the boundary of σ.

We can represent the incidence relationships of cells in X in a weighted graph B.

Definition 2.1.7 (Incidence graph)**.** *Take an abstract 2-complex X. The* incidence graph *B of X is the weighted graph with*

1. *a weight 0 node n_v for each vertex v of X,*
2. *a weight 1 node n_e for each edge $e = \{u, v\}$ of X,*
3. *a weight 2 node n_t for each triangle $t = \{u, v, w\}$ of X,*
4. *an edge between a weight 2 node n_t and weight 1 node n_e if $e \subset t$,*
5. *an edge between a weight 2 node n_t and weight 0 node n_v if $v \in t$,*
6. *an edge between a weight 1 node n_e and weight 0 node n_v if $v \in e$.*

Abusing notation, we usually write $|X|$ instead of $|X|_\Theta$ or (X, Θ), use v to denote both the abstract vertex and its embedded location $\Theta(v)$, $\overline{uv}$ to denote both the abstract edge and the embedded image $\Theta(\{\, u, v\,\})$, and $\triangle uvw$ to denote both the abstract triangle and the embedded image $\Theta(\{\, u, v, w\,\})$. Whether we are referring to an element of the abstract 2-complex or its image in $\mathbb{R}^n$ should be clear from the context.

As in Chapter 1, we use the following conventions in this chapter. Given two points $x, y \in \mathbb{R}^n$, $\|x - y\|$ is the standard Euclidean distance between x and y, for a point $x \in \mathbb{R}^n$ and a set $Y \subset \mathbb{R}^n$, we set

$$d(x, Y) := \inf_{y \in Y} \|x - y\|,$$

and for two sets $X, Y \subset \mathbb{R}^n$, we set

$$d(X, Y) := \min \left\{ \inf_{x \in X} d(x, Y), \inf_{y \in Y} d(y, X) \right\},$$
$$d_H(X, Y) := \max \left\{ \sup_{x \in X} d(x, Y), \sup_{y \in Y} d(y, X) \right\},$$

where d_H is the *Hausdorff distance*.

We also consider thickenings of a subset X: we let

$$X^\alpha := \{p \in \mathbb{R}^n \mid d(p, H) \leq \alpha\}.$$

In proofs towards the end of this chapter, we use the *weak feature size* of X to allow us to construct various isomorphisms.

Definition 2.1.8 (Weak feature size)**.** *Take $X \subset \mathbb{R}^n$ and let $d_X : \mathbb{R}^n \to \mathbb{R}$ be the distance to X function. Then the* weak feature size *of X is the infimum of all critical values of d_X.*

At various moments in the algorithm, we consider the *diameter* of a set of points.

Definition 2.1.9 (Diameter of a set of points)**.** *Let $X \subset \mathbb{R}^n$ be a finite subset of points. The* diameter *of X, $\mathcal{D}(X)$, is the maximum distance between any pair of points $x, y \in X$:*

$$\mathcal{D}(X) := \max_{x,y \in X} \|x - y\|.$$

We use $B_r(p)$ to denote the ball of radius r centred at a point $p \in \mathbb{R}^n$, by $\partial B_R(p)$ we mean the boundary of such a ball, and let

$$\mathbb{S}^k = \{x \in \mathbb{R}^n \mid \|x\| = 1\}$$

denote the standard k-sphere. We also regularly consider points in a *spherical shell*.

Definition 2.1.10. *Fix $a < b$, and let y be a point in $\mathbb{R}^n$. The* spherical shell of radii a and b *centered at p, $S_a^b(p)$ is the set*

$$\{q \in \mathbb{R}^n \mid a \leq \|q - p\| \leq b\} .$$

We consider dihedral angles between two half-planes.

Definition 2.1.11. *Let H_1, H_2 be two half-planes with a common boundary line L. Then, the* dihedral angle α *between H_1 and H_2 is the angle formed by two vectors $v_1 \in H_1$ and $v_2 \in H_2$ originating from the same point $x \in L$ such that both v_1 and v_2 are perpendicular to L.*

We work with ε-samples P of the embedded 2-complex $|X|$, which are defined analogous to Definition 1.0.3.

Definition 2.1.12 (ε-sample)**.** *Let $|X| \subset \mathbb{R}^n$ be an embedded 2-complex. An* ε-sample P *of $|X|$ is a finite subset of $\mathbb{R}^n$ such that $d_H(|X|, P) \leq \varepsilon$.*

As in Chapter 1, we use the threshold graph on a set of points, recall Definition 1.2.1.

Definition 2.1.13. *Let $P \subset \mathbb{R}^N$ be a finite collection of points, and fix $r > 0$. The* graph at threshold r on P, *$\mathfrak{G}_r(P)$, is the graph with vertices $p \in P$, and edges (p, q) if $\|p - q\| \leq r$.*

The objects we consider in this chapter are 2-dimensional, and so we also use *Čech* complexes.

Definition 2.1.14 (Čech Complex)**.** *Let $P \subset \mathbb{R}^n$ be a finite set of points. The* Čech complex at scale δ, *$\check{\mathcal{C}}_\delta(P)$ is the complex with j-cells $\{v_i\}_{i=0}^j$ such that the intersection $\bigcap_{i=0}^j B_\delta(v_i)$ is non-empty.*

Now, we formalise the aim of this chapter. Given an ε-sample P of some linearly embedded 2-complex $|X|$, we want to recover the abstract structure of the 2-complex X. To do this, we need to learn the number of vertices, the number of edges, and the number of triangles, as well as the incidence relations between them. We achieve this by first deciding for each $p \in P$ if it is near a cell that is not locally maximal, or far away from all cells which are not locally maximal. This partitions P into two subsets which intuitively are P_{NLM} containing samples p near non-locally maximal cells, and P_{LM}

containing samples p only near locally maximal cells. Rigorous definitions of P_{NLM} and P_{LM} are in Definition 2.3.6. Part of this process involves approximating the local homology at each $p \in P$ using a radius r. This requires a choice of scale at which to approximate $|X|$ from P. Unlike in Chapter 1, the relationship between clusters in P_{NLM} and P_{LM} to vertices, edges and triangles is not direct. We can, however, still infer the incidence operator.

2.2. New and Improved Geometric Lemmas

As in Section 1.1, we provide some geometric lemmas as motivation for the definitions of local structures and the geometric assumptions we place on the embeddings of a 2-complex. There are two parts to the definition of the local structure of a point cloud P at a sample p: the first is a topological condition relating to the homology of the samples in a spherical shell around p, and the second relates to the geometry of these samples. The geometric lemmas in this section allow us to distinguish between points near cells that are not locally maximal and those that are only near locally maximal cells when the topological structure of P at p does not, see Section 2.3.

We begin with a helpful lemma that bounds the distance between a point in a spherical shell within ε of a ray and the point in the ray in the middle of the shell.

Lemma 2.2.1. *Let $L \subset \mathbb{R}^n$ be a ray originating at a point z, and fix*

$$R \geq 14\varepsilon > 0.$$

Let $P \subset \mathbb{R}^n$ have $d_H(P, L) \leq \varepsilon$ and take $p \in \mathbb{R}^n$ with

$$\|p - z\| \leq \frac{R}{2}.$$

Let x be the point in L with $\|x - p\| = R$. Then for all $q \in S^{R+\varepsilon}_{R-\varepsilon}(p) \cap P$

$$\|q - x\| \leq \sqrt{2}\varepsilon.$$

Proof. Consider $S^{R+\varepsilon}_{R-\varepsilon}(p) \cap L$ say C, and a point $q \in S^{R+\varepsilon}_{R-\varepsilon}(p)$ with $d(L, q) \leq \varepsilon$. Let q_L be the projection of q to L, p_L the projection of p to L.

There are two cases we need to consider,

1. $\|x - q_L\| \geq \|q_L - p_L\|$,

2. $\|x - q_L\| < \|q_L - p_L\|$.

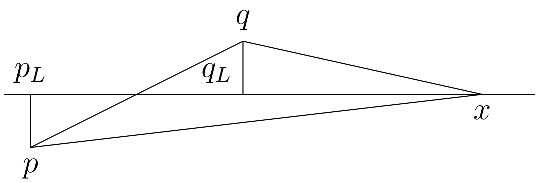


FIGURE 2.1. $\|x - q_L\| \geq \|q_L - p_L\|$

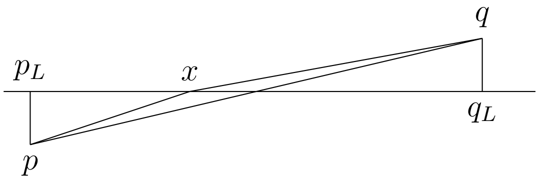


FIGURE 2.2. $\|x - q_L\| < \|q_L - p_L\|$

We begin with case 1.

We want to bound $\|x - q\|$. Note that

$$\begin{aligned}
\|q - x\|^2 &= \|q - q_L\|^2 + \|q_L - x\|^2, \\
\|q_L - x\| &= \|p_L - x\| - \|p_L - q_L\|, \\
\|p_L - q_L\|^2 &= \|q - p\|^2 - \left(\|p - p_L\| + \|q - q_L\|\right)^2, \\
\|p_L - x\|^2 &= \|x - p\|^2 - \|p_L - p\|^2.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\|q - x\|^2 \\
&= \|q - q_L\|^2 + \left(\|p_L - x\| - \|p_L - q_L\|\right)^2 \\
&= \|q - q_L\|^2 + \left(\sqrt{\|q - p\|^2 - \|p_L - p\|^2} - \sqrt{\|q - p\|^2 - \left(\|p - p_L\| + \|q - q_L\|\right)^2}\right)^2 \\
&= \|q - q_L\|^2 + \\
&\left(\sqrt{\|q - p\|^2 - \|p_L - p\|^2} - \sqrt{\|q - p\|^2 - \|p - p_L\|^2 - \left(\|q - q_L\|^2 + \|p - p_L\|\|q - q_L\|\right)}\right).
\end{aligned}$$

Let

$$\begin{aligned} A &= \|q-p\|^2 - \|p-p_L\|^2, \\ B &= \|q-q_L\|^2 + \|p-p_L\|\|q-q_L\|. \end{aligned}$$

As

$$\begin{aligned} \|q-p\| &\leq R, \\ \|p-p_L\| &\leq \frac{R}{2}, \\ \|q-q_L\| &\leq \varepsilon, \end{aligned}$$

we have

$$\begin{aligned} &A > (R-\varepsilon)^2 - \varepsilon^2 \\ &B < 3\varepsilon^2 \end{aligned}$$

and so $A > \frac{4B}{3}$. Then

$$\begin{aligned} \frac{AB}{3} &> \frac{4B^2}{9} \\ A^2 - AB &> A^2 - \frac{4AB}{3} + \frac{4B^2}{9} \\ \sqrt{A(A-B)} &> A - \frac{2B}{3} \\ -2\sqrt{A(A-B)} &< -2A + \frac{4B}{3} \\ 2A - B - 2\sqrt{A(A-B)} &< \frac{B}{3} \\ \left(\sqrt{A} - \sqrt{A-B}\right)^2 &< \frac{B}{3} \end{aligned}$$

Recall $A > \frac{4B}{3}$, thus

$$\begin{aligned} \|q-x\|^2 &= \|q-q_L\| + \left(\sqrt{A} - \sqrt{A-B}\right)^2 \\ &\leq \varepsilon^2 + \frac{B}{3} \\ &\leq 2\varepsilon \end{aligned}$$

A similar calculation in case 2 gives a smaller bound, so

$$\|q - x\| \leq \sqrt{2}\varepsilon.$$

□

Next, Lemma 2.2.2 , which motivates part 3 in Definition 2.3.4. The lemma considers the distances between triples of points in $S_{R-\varepsilon}^{R+\varepsilon}(p) \cap H^{\varepsilon}$ for some point $p \in H^{\varepsilon}$, where H^{ε} is the thickening of a plane H by ε, with $\varepsilon > 0$.

Lemma 2.2.2. *Consider an affine* 2*-hyperplane* $H \subset \mathbb{R}^n$ *and fix*

$$R \geq 14\varepsilon \geq 0.$$

Let $P \subset \mathbb{R}^n$ *be such that* $d_H(P, H) \leq \varepsilon$*, and take* p *with* $d(p, H) \leq \varepsilon$*. Then, for all* $q_1 \in S_{R-\varepsilon}^{R+\varepsilon}(p) \cap P$*, there exists* $q_2 \in S_{R-\varepsilon}^{R+\varepsilon}(p) \cap P$ *with*

$$\|q_2 - q_1\| \geq 2\sqrt{R^2 - \varepsilon^2} - (1 + \sqrt{2})\varepsilon.$$

Proof. First, let p_H be the projection of p to H, and note that $\|p_H - p\| \leq \varepsilon$. Take $q_1 \in S_{R-\varepsilon}^{R+\varepsilon}(p) \cap P$. Let x_1 be the point in $\partial B_R(p) \cap H$ closest to q_1, and q_H the projection of q_1 to H. Note that p_H, q_H, x_1 are co-linear, lying on the ray L from p_H, and $\|q_1 - q_H\| \leq \varepsilon$. By Lemma 2.2.1, $\|q_1 - x_1\| \leq \sqrt{2}\varepsilon$.

As $H \cap \partial B_R(p)$ is a circle with radius $\sqrt{R^2 - \|p_H - p\|^2}$, there is a point $x_2 \in H \cap \partial B_R(p)$ such that $\|x_2 - x_1\| = 2\sqrt{R^2 - \|p_H - p\|^2}$. As $d_H(p, H) \leq \varepsilon$, we have

$$\|x_2 - x_1\| \geq 2\sqrt{R^2 - \varepsilon^2},$$

and as $d_H(P, H) \leq \varepsilon$, there is $q_1 \in P$ with $\|q_1 - x_1\| \leq \varepsilon$. Hence

$$\|q_2 - q_1\| \geq 2\sqrt{R^2 - \varepsilon^2} - (1 + \sqrt{2})\varepsilon.$$

□

Now that we have a geometric property to test if a point p and the samples in $S_{R-\varepsilon}^{R+\varepsilon}(p)$ are from a subset of a plane. We want to understand what conditions need to be placed on points near an edge in two triangles to guarantee this property does not hold. In particular, Lemma 2.2.3 motivates part 4 of Definition 2.3.5.

For ease of reading, we let

$$\Psi(\varepsilon, R) = \arccos\left(\frac{(R+2\varepsilon)^2 + \left(\frac{3R}{2} - \varepsilon\right)^2 - \left(2\sqrt{R^2-\varepsilon^2} - \left(2+2\sqrt{2}\right)\varepsilon\right)^2}{2(R+2\varepsilon)\left(\frac{3R}{2} - \varepsilon\right)}\right).$$

The following lemma motivates the conditions we place on the dihedral angle between two triangles with a common boundary edge $\overline{uv}$ (of degree 2). This allows us to guarantee that the geometry of the samples in $S^{R+\varepsilon}_{R-\varepsilon}(p)$ for a sample p *near* $\overline{uv}$ is not the same as the geometry of samples in $S^{R+\varepsilon}_{R-\varepsilon}(p)$ when p is near a triangle but *far away* from its boundary.

Lemma 2.2.3. *Consider two affine 2-half-planes $H_1, H_2 \subset \mathbb{R}^n$ whose boundaries are equal, say L, and fix $R \geq 14\varepsilon > 0$. Let α be the dihedral angle between H_1 and H_2. Let P be a set of points such that $d_H(P, H_1 \cup H_2) \leq \varepsilon$. Further, take p such that $d(p, H_1) \leq \varepsilon$. If*

$$d(L,p) \leq \frac{R}{2} - 2\varepsilon$$

and

$$\alpha \in (0, \Psi(\varepsilon, R))$$

then there exist $q_1 \in S^{R+\varepsilon}_{R-\varepsilon}(p) \cap P$ such that for all $q_2 \in S^{R+\varepsilon}_{R-\varepsilon}(p) \cap P$

$$\|q_2 - q_1\| < 2\sqrt{R^2-\varepsilon^2} - \left(1+\sqrt{2}\right)\varepsilon.$$

Proof. First, let H_1' be the half plane containing H_1 with bounding line L' such that $D(L, L') = \varepsilon$, p_H be the projection of p onto H_1' and p_L the projection of p to L. Then take $x_1 \in H_1$ such that $\|p - x_1\| = R$ and p_H, p_L and x_1 are co-linear. Take $q_1 \in P$ with $\|q_1 - x_1\| \leq \varepsilon$, so $q_1 \in S^{R+\varepsilon}_{R-\varepsilon}(p) \cap P$.

Let q_2 be a point in $S^{R+\varepsilon}_{R-\varepsilon}(p) \cap P$. There are two cases to consider: $d(q_2, H_1') \leq \varepsilon$ and $d(q_2, H_2) \leq \varepsilon$.

If $d(q_2, H_1') \leq \varepsilon$, take $x_2 \in \partial B_R(p) \cap H_1'$ such that x_2, p_H and the projection of q_2 to H_1' are co-linear. Then by Lemma 2.2.1 $\|q_2 - x_2\| \leq \sqrt{2}\varepsilon$.

Consider the triangle formed by x_1, p_h, x_2. By assumption,

$$\|\widetilde{x} - p_H\| < \frac{R}{2} < R - 7\varepsilon,$$

$$\|x_2 - p_H\| = \|x_1 - p_H\| \leq R.$$

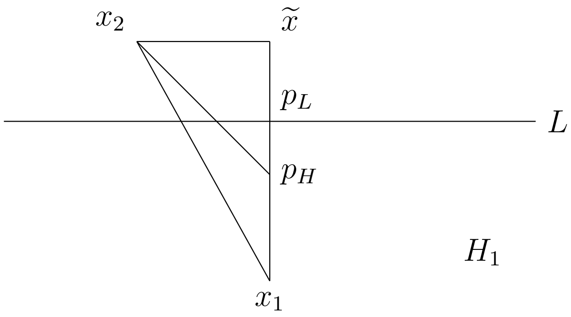


FIGURE 2.3

Let $\widehat{R} = \sqrt{R^2 - \|p_H - p\|^2}$. Then

$$\begin{aligned} \|\widetilde{x} - p_H\| &< \widehat{R} \\ \|\widetilde{x} - p_H\| &< \widehat{R} - 6\varepsilon \\ 2\widehat{R}\|\widetilde{x} - p_H\| &< 2\widehat{R}^2 - 12\widehat{R}\varepsilon \\ 2\widehat{R}^2 + 2\widehat{R}\|\widetilde{x} - p_H\| &< 4\widehat{R}^2 - 4(1+\sqrt{2})\widehat{R}\varepsilon + (1+\sqrt{2})\varepsilon \\ 2\widehat{R}^2 + 2\widehat{R}\left(\frac{\|\widetilde{x} - p_H\|}{\widehat{R}}\right) &< \left(2\widehat{R} - (1+\sqrt{2})\varepsilon\right)^2 . \end{aligned}$$

Further,

$$\begin{aligned} 2\widehat{R}^2 + 2\widehat{R}\left(\frac{\|\widetilde{x} - p_H\|}{\widehat{R}}\right) &= \|x_1 - p_H\|^2 + \|x_2 - p_H\|^2 + 2\|x_1 - p_H\|\|x_2 - p_h\| \cos \angle x_2 p_h \widetilde{x} \\ &= \|x_1 - p_H\|^2 + \|x_2 - p_H\|^2 - 2\|x_1 - p_H\|\|x_2 - p_h\| \cos \angle x_2 p_h x_1 \\ &= \|x_2 - x_1\|^2, \end{aligned}$$

so

$$\|x_2 - x_1\| < \sqrt{R^2 - \|p_H - p\|^2} - (1+\sqrt{2})\varepsilon,$$

which implies

$$\|q_2 - q_1\| < 2\sqrt{R^2 - \varepsilon^2} - (2 + 2\sqrt{2})\varepsilon.$$

Now assume $d(q_2, H_2) \leq \varepsilon$. Let H_2' be the half-plane which contains H_2 and has boundary L' with $d(L, L') = \varepsilon$. As $d(q_2, H_2) \leq \varepsilon$, then there is $x_2 \in \partial B_R(p) \cap H_2'$ with $\|q_2 - x_2\| \leq \sqrt{2}\varepsilon$. Hence,

$$\|x_1 - x_2\| \geq 2\sqrt{R^2 - \varepsilon^2} - (2 + 2\sqrt{2})\varepsilon.$$

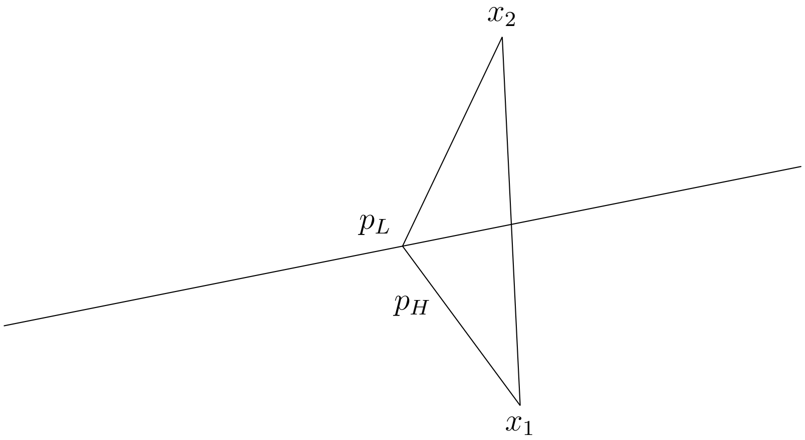


FIGURE 2.4. $d(q_2, H_2) \leq \varepsilon$

If $x_2 \in H_2' \setminus H_2$, then by a similar argument to above,

$$\|x_1 - x_2\| \geq 2\sqrt{R^2 - \varepsilon^2} - (2 + 2\sqrt{2})\varepsilon.$$

If $x_2 \in H_2 \subsetneq H_2'$, by the cosine rule we have

$$\|x_2 - x_1\|^2 = \|x_2 - p_L\|^2 + \|x_1 - p_L\|^2 - 2\|x_2 - p_L\|\|x_1 - p_L\| \cos \angle x_1 p_L x_2.$$

Note $\|x_1 - p_L\| = \|x_1 - p_H\| + \|p_H - p_L\|$, and $\|x_2 - x_1\|$ is bounded above by the case when

$$\begin{aligned} \angle x_1 p_L x_2 &= \alpha, \\ \|x_2 - p_L\| &= R + 2\varepsilon, \\ \|x_1 - p_L\| &= \|x_1 - p_H\| + \|p_H - p_L\| = \frac{3R}{2} + \varepsilon. \end{aligned}$$

Hence, we have

$$\|x_2 - x_1\| < (R + 2\varepsilon)^2 + \left(\frac{3R}{2} + \varepsilon\right)^2 - (R + 2\varepsilon)\left(\frac{3R}{2} + \varepsilon\right) \cos \alpha.$$

By assumption, $\alpha \in (0, \Psi(\varepsilon, R))$, and so

$$\|x_2 - x_1\| < 2\sqrt{R^2 - \varepsilon^2} - \left(2 + 2\sqrt{2}\right)\varepsilon,$$

which implies that

$$\|q_2 - q_1\| < 2\sqrt{R^2 - \varepsilon^2} - \left(1 + \sqrt{2}\right)\varepsilon.$$

Hence, there is a $q_1 \in S^{R+\varepsilon}_{R-\varepsilon}(p) \cap P$ such that for all $q_2 \in S^{R+\varepsilon}_{R-\varepsilon}(p) \cap P$

$$\|q_2 - q_1\| < 2\sqrt{R^2 - \varepsilon^2} - \left(1 + \sqrt{2}\right) \varepsilon.$$

□

Next, we investigate the geometry of points near a ray and half-plane, to develop a test for points near not locally maximal cells.

There are several local structures that have the same topological structure: they consist of two connected components with no 1-cycles. In Chapter 1, we used the angle between the centroids of the two connected components to distinguish between points near a degree 2 vertex and points near the *interior* of an edge. Unfortunately, this test is not sufficient after introducing triangles. If we first check for the presence of triangles, we can again use the inner-product test. To test for the presence of triangles, we examine the diameters of the two connected components.

So, we first bound the diameter of a set of samples only near a line.

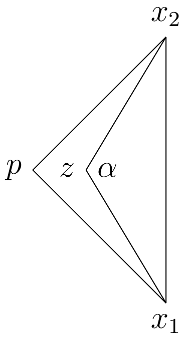
Lemma 2.2.4 (Diameter of points near ray)**.** *Let $L \subset \mathbb{R}^n$ be a ray originating at a point z, and fix $R > 14\varepsilon > 0$. Let $P \subset \mathbb{R}^n$ have $d_H(P, L) \leq \varepsilon$ and take $p \in \mathbb{R}^n$ with $d(L, p) \leq \varepsilon$ and $\|p - z\| \leq \frac{R-\varepsilon}{2}$. Then $\left(S^{R+\varepsilon}_{R-\varepsilon}(p) \cap P\right)^{\frac{3\varepsilon}{2}}$ has 1 connected component c, and the diameter is less than $2\sqrt{2}\varepsilon$.*

Proof. By Lemma 2.2.1, every $q \in S^{R+\varepsilon}_{R-\varepsilon}(p) \cap P$ is with in $\sqrt{2}\varepsilon$ of the point x in L with $\|x - p\| = R$. Hence, $\left(S^{R+\varepsilon}_{R-\varepsilon}(p) \cap P\right)^{\frac{3\varepsilon}{2}}$ consists of a single connected component and it has diameter less than $2\sqrt{2}\varepsilon$. □

The previous lemma bounds the diameter of a connected component containing points with ε of an edge, that are within $S^{R+\varepsilon}_{R-\varepsilon}(p)$ for a sample p near a vertex in the boundary of this edge. We need to guarantee that if p is near the interior of an edge, it does not fail the diameter test. To ensure this, we obtain the following as a corollary of Lemma 2.2.4.

Corollary 2.2.5. *Let $L \subset \mathbb{R}^n$ be a line, and fix $R > 3\varepsilon > 0$. Let $P \subset \mathbb{R}^n$ have $d_H(P, L) \leq \varepsilon$ and take $p \in \mathbb{R}^n$ with $d(L, p) \leq \varepsilon$ and*

$$\|p - z\| \leq \frac{R - \varepsilon}{2}.$$



Then $\left(S_{R-\varepsilon}^{R+\varepsilon}(p) \cap P\right)^{\frac{3\varepsilon}{2}}$ has 2 connected components c_1, c_2, and their diameters are less than $2\sqrt{2}\varepsilon$.

Proof. First note that $S_{R-\varepsilon}^{R+\varepsilon}(p) \cap L$ consists of two connected components, C_1, C_2, and the distance between them is $R - \varepsilon$. Hence, we can apply Lemma 2.2.4, to C_1 and C_2 individually, obtaining a connected component for each, say c_1 and c_2. Further, the diameters of c_1 and c_2 are less than $2\sqrt{2}\varepsilon$. □

The following lemma guarantees that if there are samples in $S_{R-\varepsilon}^{R+\varepsilon}(p)$ that are within ε of a triangle, the diameter test fails.

Lemma 2.2.6. *Let $L_1, L_2 \subset \mathbb{R}^n$ be two rays originating at the same point z with the angle α between in the interval*

$$\left[\frac{\pi}{6}, \pi\right),$$

and fix $R \geq 14\varepsilon > 0$. Let T be the set between L_1 and L_2. Take $p \in \mathbb{R}^n$ with $d(T, p) \leq \varepsilon$ and $\|p - x\| \leq \frac{R-\varepsilon}{2}$, and $P \subset \mathbb{R}^n$ with $d_H(P, T) \leq \varepsilon$. Then, there exist points q_1, q_2 in P with $\|q_1 - p\|, \|q_2 - p\| \in [R - \varepsilon, R + \varepsilon]$ such that $\|q_1 - q_2\| > 2\sqrt{2}\varepsilon$, and q_1, q_2 are path connected. Furthermore, the connected component containing q_1 and q_2 has diameter bigger than $2\sqrt{2}\varepsilon$.

Proof. As $\|p - z\| \leq \frac{R-\varepsilon}{2}$, the intersection $S_{R-\varepsilon}^{R+\varepsilon}(p) \cap T$ is not empty, connected, and $\mathcal{H}_1\left(S_{R-\varepsilon}^{R+\varepsilon}(p) \cap T\right) = 0$. Further, the intersections $S_{R-\varepsilon}^{R+\varepsilon}(p) \cap L_1$ and $S_{R-\varepsilon}^{R+\varepsilon}(p) \cap L_2$ are also connected.

Now, let x_1 be the point on L_1 with $\|q_1 - p\| = R$ and let x_2 be the point on L_2 with $\|x_2 - p\| = R$. As $S_{R-\varepsilon}^{R+\varepsilon}(p) \cap T$ is path connected, x_1 and x_2 are path connected in T.

Consider the triangle $\triangle x_1 p x_2$, we have

$$\begin{aligned}
\|x_1 - x_2\|^2 &= \|x_1 - z\|^2 + \|x_2 - z\|^2 - 2\|x_1 - z\|\|q_2 - z\| \cos\alpha \\
&\geq \left(R - \frac{R-\varepsilon}{2}\right)^2 + \left(R - \frac{R-\varepsilon}{2}\right)^2 - 2\left(R - \frac{R-\varepsilon}{2}\right)^2 \cos\alpha \\
&= 2\left(\frac{R+\varepsilon}{2}\right)^2 (1 - \cos\alpha).
\end{aligned}$$

Now, as $d_H(P,T) \leq \varepsilon$, there are points $q_1, q_2 \in P$ with

$$\|q_1 - x_1\|, \|q_2 - x_2\| \leq \varepsilon.$$

Then by the triangle inequality

$$\begin{aligned}
\|q_1 - q_2\|^2 &= 2\left(\frac{R+\varepsilon}{2}\right)^2 (1 - \cos\alpha) - 2\varepsilon \\
&> 2\sqrt{2}\varepsilon, \text{ as } \alpha \in \left[\frac{\pi}{6}, \pi\right).
\end{aligned}$$

□

2.3. Local Structures

To identify the abstract structure of the 2-complex, the algorithm in Section 2.4 first partitions the sample P into sets P_{LM}, containing samples that are only near locally maximal cells, and P_{NLM}, containing samples near cells that are not locally maximal. The decision tree for if a point is in P_{NLM} or P_{LM} is summarised in Figure 2.5. After this, we further partition P_{LM} and P_{NLM} to infer the number of cells and their dimensions, as well as the incidence operator.

Take an embedded 2-complex $|X| \subset \mathbb{R}^n$, fix (an appropriate) $0 < \varepsilon \leq R$ and take $p \in \mathbb{R}^n$ with $d(|X|, p) \leq \varepsilon$. Consider the topological and geometric structure of $|X|$ in a neighbourhood of p, beginning with $B_R(p) \cap |X|$. If $B_R(p) \cap |X|$ is disconnected, we restrict to the connected component C_p containing $\mathrm{proj}_{|X|}(p)$. Then, we consider $\partial B_R(p) \cap C_p$. Let $\mathrm{proj}_{|X|}(p)$ be the projection of p to $|X|$, and let σ_p be the cell containing $\mathrm{proj}_{|X|}(p)$. If σ_p

is locally maximal and $d(|\partial\sigma_p|, p) > R$, then $\partial B_R(p) \cap C_p$ has one of the following structures:

1. $\partial B_R(p) \cap C_p$ is empty, in which case σ_p is a locally maximal vertex,
2. $\partial B_R(p) \cap C_p$ is a pair of antipodal points, in which case σ_p is a locally maximal 1-cell,
3. $\partial B_R(p) \cap C_p$ is homotopic to $\mathcal{S}^1$ lying in a plane, in which case σ_p is a 2-cell.

The above structures consist of two parts: we examine the topological structure of $\partial B_R(p) \cap C_p$, and then look at its geometry. If p is within R of some cell τ_p (possibly $\tau_p = \sigma_p$) which is not locally maximal, then either the topological structure or the geometric structure is not one of the above cases. As such, we use a two-step process to decide if a given sample p is within R of some not locally maximal cell τ_p: first, we examine the topological structure of $\partial B_R(p) \cap C_p$ by looking at its homology, and then if necessary, we consider its geometric structure. We let

$$\mathcal{H}_\bullet(p) := H_\bullet\left(\partial B_R(p) \cap C_p\right).$$

As we are restricting ourselves to 2-complexes, we focus on $\mathcal{H}_0(p)$ and $\mathcal{H}_1(p)$.

Definition 2.3.1 (Local homology signature)**.** *Let* $|X| \subset \mathbb{R}^n$ *be an embedded* 2*-complex, and fix* $R > \varepsilon > 0$*. Take a point* $p \in \mathbb{R}^n$ *with* $d(p, |X|) \leq \varepsilon$*. The* local homology signature *of* $|X|$ *at* p *is*

$$\mathrm{Sig}(p) := \left(|\mathcal{H}_0(p)|, |\mathcal{H}_1(p)|\right).$$

In the above cases, the local homology signature of $|X|$ at p is as follows.

1. $\mathrm{Sig}(p) = (0, 0)$,
2. $\mathrm{Sig}(p) = (2, 0)$,
3. $\mathrm{Sig}(p) = (1, 1)$.

and so if $\mathrm{Sig}(p)$ is not equal to $(0, 0)$, $(2, 0)$ or $(1, 1)$, then p is within R of a cell τ_p which is not locally maximal. If $\mathrm{Sig}(p)$ is $(0, 0)$ then p is within ε of a degree 0 vertex. Unfortunately, if $\mathrm{Sig}(p)$ is either $(2, 0)$ or $(1, 1)$, we need to examine the geometric structure of $\partial B_R(p) \cap C_p$. When $\mathrm{Sig}(p) = (2, 0)$, we can distinguish between the case where σ_p is a locally maximal 1-cell and where σ_p is a vertex of degree 2 as follows: let the two points in $\partial B_R(p) \cap C_p$ be c_1 and c_2. If σ_p is a 1-cell, then $\angle c_1 p c_2 = \pi$, and other $\angle c_1 p c_2 \neq \pi$. When $\mathrm{Sig}(p) = (1, 1)$ we need to distinguish between if σ_q is a 2-cell, and if σ_p is in the boundary of 2-cells. We can do so by checking if $\partial B_R(p) \cap C_p$ is contained in a plane: if it is, then σ_p is a 2-cell, if not σ_p is either an edge or a vertex that is not locally maximal.

Recall that we are working with an ε-sample P of the embedded 2-complex $|X|$ instead of $|X|$. We want to approximate $\mathrm{Sig}(p)$ with P. As P is an ε-sample, we can approximate $\partial B_R(p) \cap C_p$ by first considering the structure of $B_{R+\varepsilon}(p) \cap P$, then the structure of $S_{R-\varepsilon}^{R+\varepsilon}(p) \cap P$. Before we define the (ε, R)-local structure of P at p (Definition 2.3.3), we need the following notation.

Definition 2.3.2. *Let $P \subset \mathbb{R}^n$ be a finite set of points. Then, $\mathrm{rk}_k^{\delta,\gamma}(P)$ is the rank of the map on the k^{th} homology groups induced by the inclusion $P^\delta \hookrightarrow P^\gamma$.*

We can now formally define the (ε, R)-local structure of P at p.

Definition 2.3.3 ((ε, R)-local homology signature)**.** *Let $P \subset \mathbb{R}^n$ be an ε-sample of an embedded 2-complex $|X|$, and fix $R \geq 14\varepsilon$. Let $C_p^{\frac{3\varepsilon}{2}}$ be samples in the same connected component of threshold graph $\mathfrak{G}_{3\varepsilon}\left(B_{R+\varepsilon}(p) \cap P\right)$ as p. The (ε, R)-*local homology signature $\mathrm{Sig}_{\varepsilon,R}(p)$ *of P at a sample p is*

$$\mathrm{Sig}_{\varepsilon,R}(p) := \left(\mathrm{rk}_0^{\frac{3\varepsilon}{2},\frac{7\varepsilon}{2}}\left(S_{R-\varepsilon}^{R+\varepsilon}(p) \cap C_p^{\frac{3\varepsilon}{2}}\right), \mathrm{rk}_1^{\frac{3\varepsilon}{2},\frac{7\varepsilon}{2}}\left(S_{R-\varepsilon}^{R+\varepsilon}(p) \cap C_p^{\frac{3\varepsilon}{2}}\right)\right).$$

We now define the types of local structures, beginning with *maximal* local structures.

Definition 2.3.4 (Maximal (ε, R)-local structure)**.** *Let P be an ε sample of a linearly embedded 2-complex $|X|$ and fix $R \geq 14\varepsilon$. Let $C_p^{\frac{3\varepsilon}{2}}$ be the set of samples in the same connected component of $\left(B_{R+\varepsilon}(p) \cap P\right)^{\frac{3\varepsilon}{2}}$ as p. We say the (ε, R)-*local structure of P at p is maximal *if any of the following hold:*

1. $\mathrm{Sig}_{\varepsilon,R}(p) = (0, 0)$*, in which case we say that the (ε, R)-*local structure of P at p is maximal of dimension 0,
2. $\mathrm{Sig}_{\varepsilon,R}(p) = (2, 0)$*, and the two connected components c_1, c_2 of $\left(S_{R-\varepsilon}^{R+\varepsilon}(p) \cap C_p^{\frac{3\varepsilon}{2}}\right)^{\frac{3\varepsilon}{2}}$ have diameters less than 5ε and mid-points q_1*

and q_2 such that

$$\langle q_1 - p, q_2 - p\rangle \leq -R^2 + 2R\varepsilon + 7\varepsilon^2,$$

in which case we say that the (ε, R)-local structure of P at p is maximal of dimension 1*,*

3. $\mathrm{Sig}_{\varepsilon,R}(p) = (1,1)$*, and for all* $q_1 \in S^{R+\varepsilon}_{R-\varepsilon}(p) \cap P, \exists q_2 \in S^{R+\varepsilon}_{R-\varepsilon}(p) \cap P$ *with*

$$\|q_2 - q_1\| < 2\sqrt{R^2 - \varepsilon^2} - \left(1 + \sqrt{2}\right)\varepsilon.$$

in which case we say that the (ε, R)-local structure of P at p is maximal of dimension 2*,*

Next, we define *not maximal* (ε, R)-local stuctures.

Definition 2.3.5 (Not maximal (ε, R)-local structure)**.** *Let P be an ε sample of a linearly embedded 2-complex $|X|$ and fix $R \geq 14\varepsilon$. Let $C_p^{\frac{3\varepsilon}{2}}$ be the set of samples in the same connected component of $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}\left(S_{R+\varepsilon}(p) \cap P\right)$ as p. We say that* the (ε, R)-local structure of P at $p \in P$ is not maximal *if any of the following hold:*

1. $\mathrm{Sig}_{\varepsilon,R}(p) = (n, 0)$ *for some* $n \in \mathbb{Z}_{\geq 0}$, $n \neq 0, 2$*,*
2. $\mathrm{Sig}_{\varepsilon,R}(p) = (1, n)$ *for some* $n \in \mathbb{Z}_{\geq 0}$, $n \neq 1$*,*
3. $\mathrm{Sig}_{\varepsilon,R}(p) = (2, 0)$ *and letting two connected components of*

$$\left(S^{R+\varepsilon}_{R-\varepsilon}(p) \cap C_p^{\frac{3\varepsilon}{2}}\right)^{\frac{3\varepsilon}{2}}$$

be c_1, c_2*, either* $\max\{\mathcal{D}(c_1), \mathcal{D}(c_2)\} \leq 2\sqrt{2}\varepsilon$ *and letting mid-points of* c_1, c_2 *be* q_1, q_2

$$\langle q_1 - p, q_2 - p\rangle > -R^2 + 2R\varepsilon + 7\varepsilon^2,$$

4. $\mathrm{Sig}_{\varepsilon,R}(p) = (1, 1)$ *and there exists* $q_1 \in P \cap S^{R+\varepsilon}_{R-\varepsilon}$ *such that for all* $q_2 \in P \cap S^{R+\varepsilon}_{R-\varepsilon}$

$$\|q_2 - q_1\| < 2\sqrt{R^2 - \varepsilon^2} - \left(1 + \sqrt{2}\right)\varepsilon.$$

Having defined the two classes of (ε, R)-local structures, we can define our initial partition.

Definition 2.3.6 (P_{LM} and P_{NLM}). *Let P be an ε-sample of an embedded 2-complex $|X|$. We partition P into two sets P_{LM} and P_{NLM} defined as*

$$P_{LM} := \{p \in P \mid \text{ the } (\varepsilon, R)\text{-local structure at of } P \text{ at } p \text{ is maximal.}\}$$
$$P_{NLM} := \{p \in P \mid \text{ the } (\varepsilon, R)\text{-local structure of } P \text{ at } p \text{ is not maximal.}\}$$

Remark 2.3.7. For all $p \in P$, P either has maximal (ε, R)-local structure at $p \in P$ or it does not. Hence, the partitioning of P into P_{LM} and P_{NLM} defined in Definition 2.3.6 is disjoint.

Recall that the samples we are working with can contain noise, and we use the homology of $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}\left(S^{R+\varepsilon}_{R-\varepsilon}(p) \cap C_p^{\frac{3\varepsilon}{2}}\right)$ in the definition of (ε, R)-local structure. Hence, we place assumptions on $|X|$ to ensure that we correctly detect when samples are near cells that are not locally maximal. As in Chapter 1, we place assumptions on the distances between any two vertices u and v, the distance between an edge $\overline{uw}$ and a vertex $v \neq u, w$, the angle between any pair of edges with a common boundary vertex. Additionally, we place assumptions on the dihedral angle between any two 2-cells with common boundary components. So that we can infer the incidence operator, we will require an upper bound on the relationship between R and ε, and so we also restrict our choice of R in terms of ε. We use the following notation in the decision flow chart (Figure 2.5):

$$\beta = -R^2 + 2R\varepsilon + 7\varepsilon^2,$$
$$\gamma = 2\sqrt{R^2 - \varepsilon^2} - \left(1 + \sqrt{2}\right)\varepsilon.$$

To increase the readability of this chapter, we define the following functions.

Definition 2.3.8. *Fix $R > 14\varepsilon > 0$. We define the following functions:*

1.

$$\begin{aligned}\Psi_1(\varepsilon, R) &= \arccos\left(\frac{\left(\frac{R}{2} - \varepsilon\right)^2 - 18\varepsilon^2}{\left(\frac{R}{2} - \varepsilon\right)^2}\right) \\ &\geq \arccos\left(\frac{(R - \varepsilon)^2 - 18\varepsilon^2}{(R - \varepsilon)^2}\right) \\ &\quad + 2\arcsin\left(\frac{2\varepsilon}{(R - \varepsilon)}\right)\end{aligned}$$

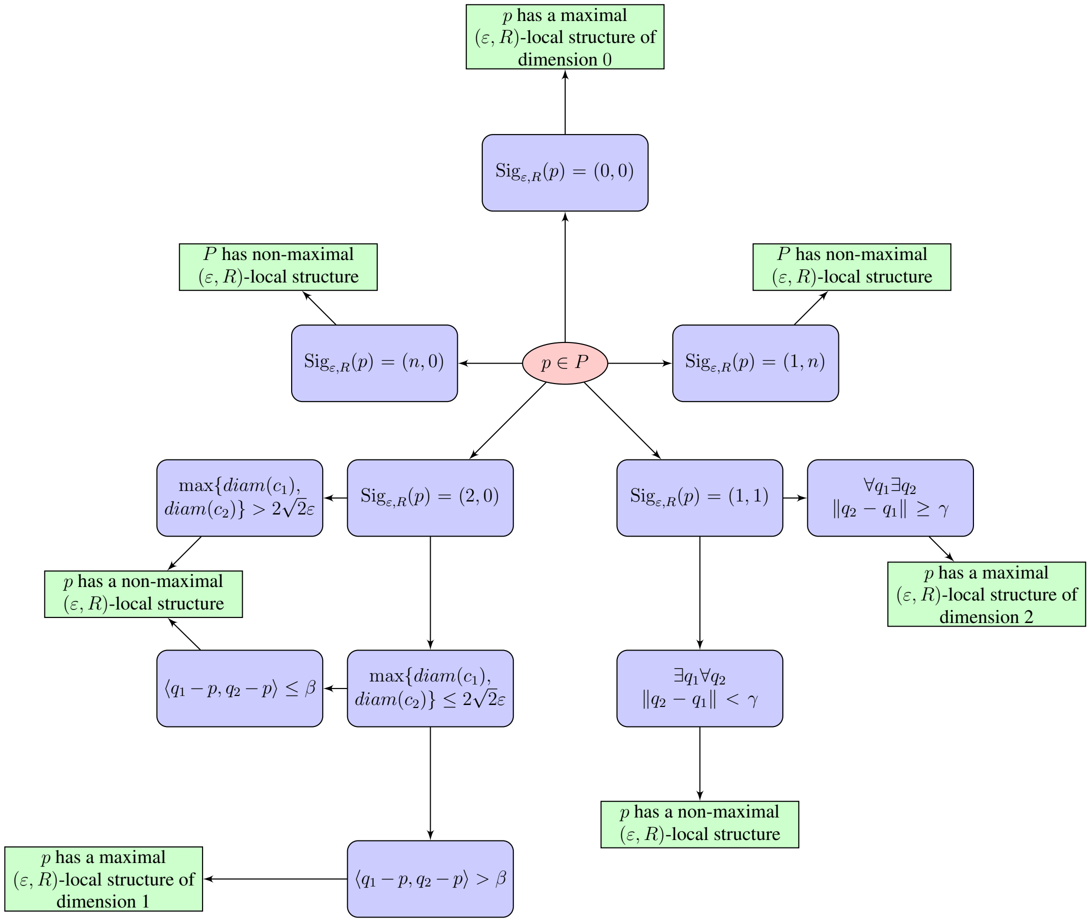

p has a maximal (ε, R)-local structure of dimension 0
$\mathrm{Sig}_{\varepsilon,R}(p) = (0,0)$
P has non-maximal (ε, R)-local structure
P has non-maximal (ε, R)-local structure
$\mathrm{Sig}_{\varepsilon,R}(p) = (n,0)$
$p \in P$
$\mathrm{Sig}_{\varepsilon,R}(p) = (1,n)$
$\max\{diam(c_1), diam(c_2)\} > 2\sqrt{2}\varepsilon$
$\mathrm{Sig}_{\varepsilon,R}(p) = (2,0)$
$\mathrm{Sig}_{\varepsilon,R}(p) = (1,1)$
$\forall q_1 \exists q_2 \; \|q_2 - q_1\| \geq \gamma$
p has a non-maximal (ε, R)-local structure
p has a maximal (ε, R)-local structure of dimension 2
$\langle q_1 - p, q_2 - p\rangle \leq \beta$
$\max\{diam(c_1), diam(c_2)\} \leq 2\sqrt{2}\varepsilon$
$\exists q_1 \forall q_2 \; \|q_2 - q_1\| < \gamma$
p has a non-maximal (ε, R)-local structure
p has a maximal (ε, R)-local structure of dimension 1
$\langle q_1 - p, q_2 - p\rangle > \beta$

FIGURE 2.5. Flow chart for determining if the (ε, R)-local structure of P at p is maximal or not. If maximal, what the dimension is.

2.

$$\Psi_2(\varepsilon, R) = \pi - \arctan\left(\frac{R + 3\varepsilon}{6\varepsilon}\right) + \arcsin\left(\frac{R^2 - 4R\varepsilon - 9\varepsilon^2}{(R+\varepsilon)\sqrt{R^2 + 6R\varepsilon + 34\varepsilon^2}}\right)$$

3.

$$\Psi_3(\varepsilon, R) = \arccos\left(\frac{(R+2\varepsilon)^2 + \left(\frac{3R}{2} - \varepsilon\right)^2 - \left(2\sqrt{R^2 - \varepsilon^2} - \left(2 + 2\sqrt{2}\right)\varepsilon\right)^2}{2(R+2\varepsilon)\left(\frac{3R}{2} - \varepsilon\right)}\right)$$

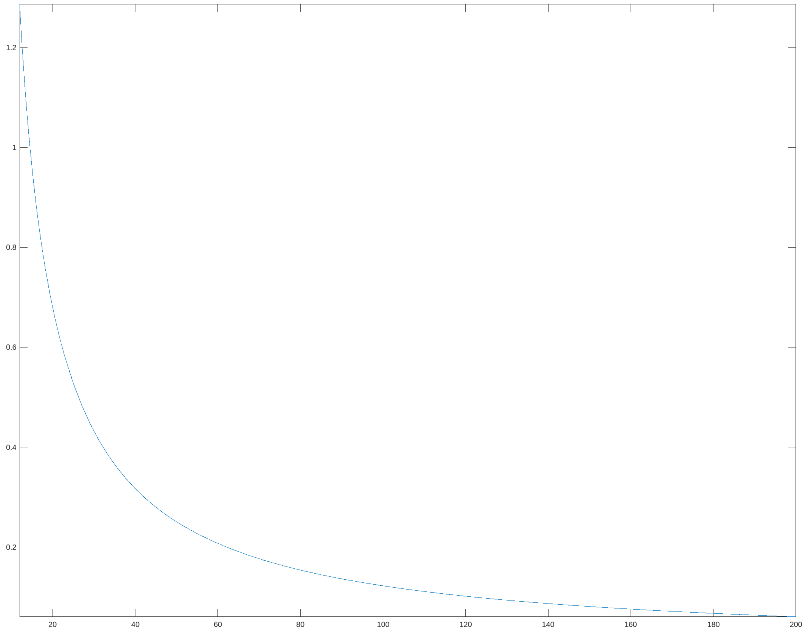


FIGURE 2.6. Graph of $\Psi_1\left(1, \frac{R}{\varepsilon}\right)$.

To improve intuition of these functions, Figures 2.6 to 2.8 provide graphs of them. Note they are effectively a function of $\frac{R}{\varepsilon}$ as they are invariant to scaling both R and ε by the same amount.

We now state the assumptions we place on $|X|$.

Assumption 2. *Fix $R \geq 14\varepsilon > 0$. We restrict to embedded 2-complexes $|X|= (X, \pi)$ which satisfy the following.*

1. *For all vertices u, v,*

$$\|u - v\| > 6(R + \varepsilon).$$

2. *For a vertex v and edge $\overline{uw}$ with $v \neq u, w$,*

$$d(\overline{uw}, v) > 6(R + \varepsilon)\varepsilon.$$

3. *For a vertex v and a triangle $\triangle uwx$ with $v \neq u, w, x$,*

$$d(\triangle uwx, v) > 6(R + \varepsilon).$$

4. *For an edge $\overline{uv}$ and a triangle $\triangle wxy$ with $v, u \neq w, x, y$,*

$$d(\triangle wxy, \overline{uv}) > 6(R + \varepsilon).$$

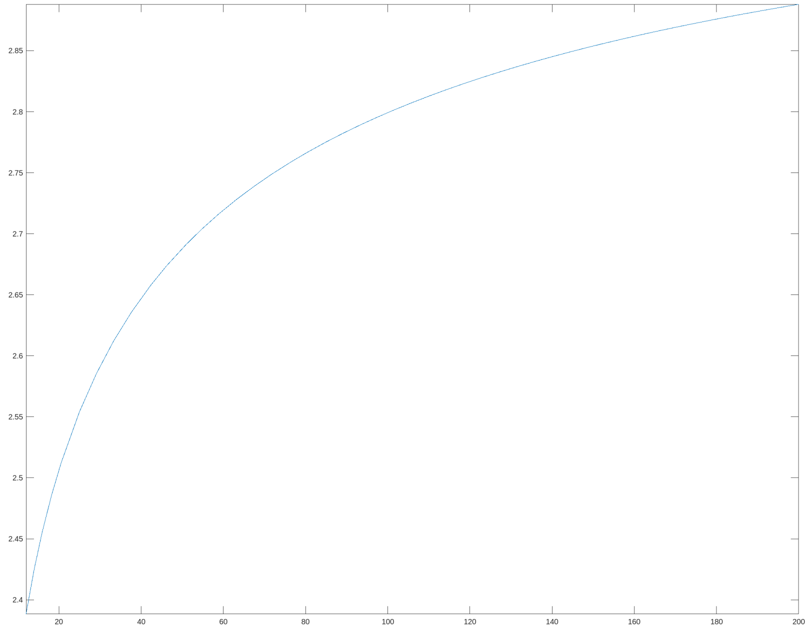


FIGURE 2.7. Graph of $\Psi_2\left(1, \frac{R}{\varepsilon}\right)$.

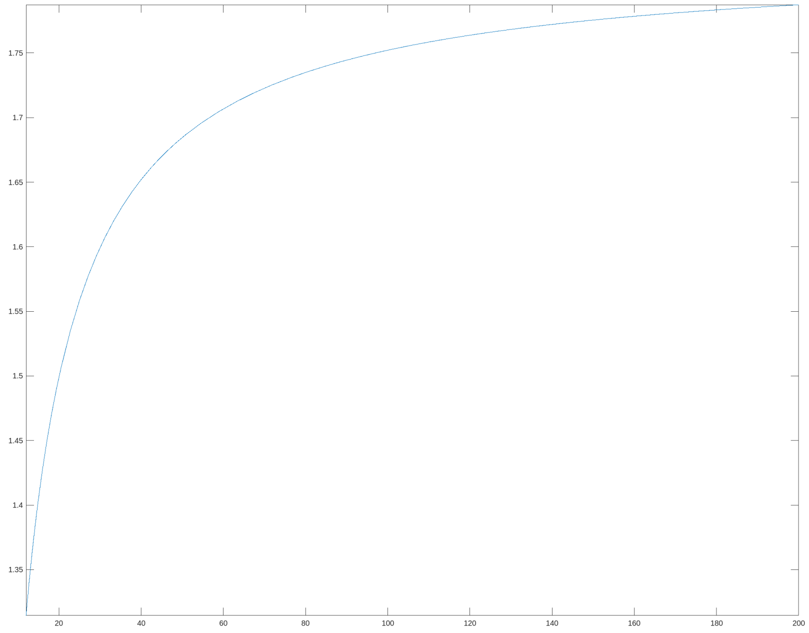


FIGURE 2.8. Graph of $\Psi_3\left(1, \frac{R}{\varepsilon}\right)$.

5. *For any triangle* $\triangle uvw$,

$$\angle uvw, \angle vwu, \angle wuv \geq \frac{\pi}{6}.$$

6. *For any pair of edges $\overline{uv}, \overline{xy}$ with no common vertex,*
$$d(\overline{uv}, \overline{xy}) > 6(R+\varepsilon).$$
7. *For any triangles $\triangle uwv, \triangle xyz$,*
$$d(\triangle uwv, \triangle xyz) > 6(R+\varepsilon).$$
8. *For any pair of edges $\overline{uv}, \overline{wv}$,*
$$\angle uvw \geq \Psi_1(\varepsilon, R).$$
9. *For all degree 2 vertices v with edges $\overline{uv}, \overline{wv}$ and no triangle $\triangle uvw$,*
$$\angle uvw \leq \Psi_2(\varepsilon, R).$$
10. *For any pair of triangles $\triangle uvw_1, \triangle uvw_2$, the dihedral angle between them is bounded below by $\Psi_1(\varepsilon, R)$.*
11. *For any pair of triangles $\triangle uvw_1, \triangle uvw_2$, with $\overline{uv}$ of degree 2, the dihedral angle between them is bounded above by $\Psi_2(\varepsilon, R)$.*
12. *For any triangle $\triangle wwvw_2$ and edge $\overline{uv}$ the angle between $\overline{uv}$ and and ray L in $\triangle w_1vw_2$ at v is bounded below by $\Psi_1(\varepsilon, R)$ and the radius of the largest circle inscribed by $\triangle uvw$ is at least $2R+3\varepsilon$.*
13. *For any vertex v such that*
$$|H_0\left(B_R(v)\cap|X|\right)|= 1, \textit{ and } |H_1\left(B_R(v)\cap|X|\right)|= 1,$$
the angle between any two rays $L_1, L_2 \in |X|$ at v is bounded above $\Psi_3(\varepsilon, R)$.

Remark 2.3.9. The reasons behind some of the conditions in Assumption 2 are relatively clear, while others are a bit more obscure. In particular, the roles of conditions 11 and 12 are not immediately clear. Condition 12 allows us to detect the vertex v in our algorithms. In particular, it is used in Proposition 2.3.14 show that we obtain $\mathrm{Sig}_{\varepsilon,R} = (n, \bullet),\ \ n \geq 2$. Condition 13 allows us to detect which topologically looks similar to an edge of degree 2 or a triangle, and so we place restrictions on the formation of the *cone*, potentially with *fins*, so that we can detect the vertex (Proposition 2.3.14). This condition is equivalent to bounding the angle at v of the convex hull which contains the triangles with vertex v.

The following Propositions provide us with 'regions' near locally maximal i-cells σ (for $i = 0,\ 1,\ 2$), where we can guarantee that at any sample in this region, the (ε, R)-local structure of P at p is maximal of dimension i.

We begin with the region around a locally maximal vertex.

Proposition 2.3.10. *Let v be a vertex of $|X| \subset \mathbb{R}^n$, which is locally maximal, and let P be an ε-sample of $|X|$. Then, for all $p \in P$ with $\|p - v\| \leq 4\varepsilon$, the (ε, R)-local structure of P at p is maximal of dimension* 0.

Proof. As v is locally maximal, it is not in the boundary of any other cell, and from Assumption 2 for all vertices $u \neq v$, $\|u - v\| > 6(R + \varepsilon)$, for all edges $\overline{uw}$ with $v \neq u, w$,

$$d(\overline{uv}, v) > 6(R + \varepsilon),$$

and for all triangles $\triangle uwx$ with $v \neq u, w, x$,

$$d(\triangle uwx, v) > 6(R + \varepsilon).$$

Hence, any sample $p \in P$ within 4ε of v is within ε of v. Thus, $(B_{R+\varepsilon}(p) \cap P)^{\frac{3\varepsilon}{2}}$ consists of a single connected component, and $S^{R+\varepsilon}_{R-\varepsilon}(p) \cap P = \emptyset$.

Thus, $S^{R+\varepsilon}_{R-\varepsilon}(p) \cap P$, $\mathrm{Sig}_{\varepsilon,R}(p) = (0, 0)$, and the (ε, R)-local structure of P at p is maximal of dimension 0. □

Next, we bound the region near a locally maximal edge.

Proposition 2.3.11. *Let $\overline{uv}$ be an edge of $|X| \subset \mathbb{R}^n$, which is locally maximal, and let P be an ε-sample of $|X|$. Then, for all $p \in P$ with $d(\overline{uv}, p) \leq \varepsilon$, and $\|p - u\|, \|p - v\| \geq \frac{3R}{2} + \varepsilon$, the (ε, R)-local structure of P at p is maximal of dimension* 1.

Proof. By Assumption 2, for any vertex $w \neq u, v$

$$d(\overline{uv}, w) > 6(R + \varepsilon),$$

for any edge $\overline{wx}$, with $w, x \neq u, v$,

$$d(\overline{uv}, \overline{wx}) > 6(R + \varepsilon),$$

for any triangle $\triangle wxy$, with $w, x, y \neq u, v$,

$$d(\triangle wxy, \overline{uv}) > 6(R+\varepsilon),$$

and so the connected component $C_p^{\frac{3\varepsilon}{2}}$ of $(B_{R+\varepsilon}(p) \cap P)^{\frac{3\varepsilon}{2}}$ which contains p, contains only points $q \in P$ with $d(q, \overline{uv}) \leq \varepsilon$.

Hence, $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}\left(S_{R-\varepsilon}^{R+\varepsilon}(p) \cap C_p^{\frac{3\varepsilon}{2}}\right)$ consists of two connected components, c_1 and c_2. By Lemma 2.2.4, the diameters of c_1 and c_2 are less than 5ε . Let x_1 and x_2 be the centroids of c_1 and c_2. Then, applying Lemma 1.1.1,

$$\langle x_1 - p, x_2 - p \rangle \leq -R^2 + 2R\varepsilon + 7\varepsilon^2,$$

so the (ε, R)-local structure of P at p is maximal of dimension 1. □

Finally, we bound the region near (locally maximal) triangles.

Proposition 2.3.12. *Let $\triangle uvw$ be an triangle of $|X| \subset \mathbb{R}^n$, and let P be an ε-sample of $|X|$. Then, for all $p \in P$ with $d(\triangle uvw, p) \leq \varepsilon$, and $d(\partial\triangle uvw, p) \geq \frac{3R}{2} + \varepsilon$, the (ε, R)-local structure of P at p is maximal of dimension 2.*

Proof. From Assumption 2, for all triangles $\triangle xyz$, with $x, y, z \neq u, v, w$,

$$d(\triangle uwv, \triangle xyz) > 6(R+\varepsilon),$$

and hence the connected component $C_p^{\frac{3\varepsilon}{2}}$ of $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(B_{R+\varepsilon}(p) \cap P)$ containing p, consists only of samples $q \in P$ with $d(q, \triangle uvw) \leq \varepsilon$, as the angle between triangles is bounded below (Assumption 2).

First, we need to show that $\mathrm{Sig}_{\varepsilon,R}(p) = (1,1)$, after which Lemma 2.2.2 implies that for all $q_1 \in S_{R-\varepsilon}^{R+\varepsilon}(p) \cap P$, there exists $q_2 \in S_{R-\varepsilon}^{R+\varepsilon}(p) \cap P$ such that

$$\|q_2 - q_1\| \geq 2\sqrt{R^2 - \varepsilon^2} - (1+\sqrt{2})\varepsilon.$$

As $d(\partial\triangle uvw, p) > \frac{3R}{2} + \varepsilon$, we have the following inclusions

$$\begin{aligned} S^{R+\varepsilon}_{R-\varepsilon}(p) \cap \triangle uvw &\hookrightarrow \left(S^{R+\varepsilon}_{R-\varepsilon}(p) \cap P\right)^{\frac{3\varepsilon}{2}} \\ &\hookrightarrow \left(S^{R+\varepsilon}_{R-\varepsilon}(p) \cap \triangle uvw\right)^{\frac{5\varepsilon}{2}} \\ &\hookrightarrow \left(S^{R+\varepsilon}_{R-\varepsilon}(p) \cap P\right)^{\frac{7\varepsilon}{2}} \\ &\hookrightarrow \left(S^{R+\varepsilon}_{R-\varepsilon}(p) \cap \triangle uvw\right)^{\frac{9\varepsilon}{2}} . \end{aligned}$$

By the bounds in Assumption 2 on the distances between a triangle and cells not in its boundary, the weak feature size of $S^{R+\varepsilon}_{R-\varepsilon}(p) \cap \triangle uvw$ is greater than 5ε, and so the inclusion maps induce isomorphisms

$$H_\bullet\left(S^{R+\varepsilon}_{R-\varepsilon}(p) \cap \triangle uvw\right) \cong H_\bullet\left(\left(S^{R+\varepsilon}_{R-\varepsilon}(p) \cap \triangle uvw\right)^{\frac{5\varepsilon}{2}}\right) \cong H_\bullet\left(\left(S^{R+\varepsilon}_{R-\varepsilon}(p) \cap \triangle uvw\right)^{\frac{9\varepsilon}{2}}\right).$$

The above homology factors through $\left(S^{R+\varepsilon}_{R-\varepsilon}(p) \cap P\right)^{\frac{3\varepsilon}{2}}$ and $\left(S^{R+\varepsilon}_{R-\varepsilon}(p) \cap P\right)^{\frac{7\varepsilon}{2}}$ so we have

$$\mathrm{rk}_\bullet^{\frac{3\varepsilon}{2},\frac{5\varepsilon}{2}}\left(S^{R+\varepsilon}_{R-\varepsilon}(p) \cap P\right) = \left|H_\bullet\left(S^{R+\varepsilon}_{R-\varepsilon}(p) \cap \triangle uvw\right)\right|,$$

and as

$$|H_0\left(S^{R+\varepsilon}_{R-\varepsilon}(p) \cap \triangle uvw\right)| = 1, |H_1\left(S^{R+\varepsilon}_{R-\varepsilon}(p) \cap \triangle uvw\right)| = 1,$$

it follows that $\mathrm{Sig}_{\varepsilon,R}(p) = (1,1)$. Now we apply Lemma 2.2.2 and conclude that the (ε, R)-local structure of P at p is maximal of dimension 2. □

Now, we obtain the regions around not locally maximal i-cells σ ($i = 0, 1$) in which we can guarantee that the (ε, R)-local structure of P at a sample p in this region is not locally maximal. Again, we begin with non-locally maximal vertices.

Remark 2.3.13. As we have restricted our considerations to 2-complexes, every triangle σ is locally maximal; hence, we need only to consider vertices and edges that are not locally maximal.

Proposition 2.3.14. *Let v be a vertex of $|X| \subset \mathbb{R}^n$, which is not locally maximal, and let P be an ε-sample of $|X|$. Then, for all $p \in P$ with*

$$\|p - v\| \leq \frac{R}{2} - 2\varepsilon,$$

the (ε, R)-local structure of P at p is not maximal.

Proof. There are several cases we need to consider, which we can classify by the homology of $\partial B_R(v) \cap |X|$:

1. $|H_0 (\partial B_R(v) \cap |X|)| = n$, $|H_1 (\partial B_R(v) \cap |X|)| = 0$, $n \neq 2$,
2. $|H_0 (\partial B_R(v) \cap |X|)| = 2$, $|H_1 (\partial B_R(v) \cap |X|)| = 0$,
3. $|H_0 (\partial B_R(v) \cap |X|)| = 1$, $|H_1 (\partial B_R(v) \cap |X|)| = 1$,
4. $|H_0 (\partial B_R(v) \cap |X|)| = 1$, $|H_1 (\partial B_R(v) \cap |X|)| = n$, $n \geq 2$.

In each of these cases, the following argument holds. Let C_p be the connected component of $B_{R+\varepsilon}(p) \cap |X|$ which contains the projection of p to $|X|$, and let $C_p^{\frac{3\varepsilon}{2}}$ be the connected component of $\left(S_{R-\varepsilon}^{R+\varepsilon}(p) \cap P\right)^{\frac{3\varepsilon}{2}}$. As P is a ε-sample of $|X|$, we have the following inclusions

$$\begin{aligned} S_{R-\varepsilon}^{R+\varepsilon}(p) \cap \triangle uvw &\hookrightarrow \left(S_{R-\varepsilon}^{R+\varepsilon}(p) \cap P\right)^{\frac{3\varepsilon}{2}} \\ &\hookrightarrow \left(S_{R-\varepsilon}^{R+\varepsilon}(p) \cap \triangle uvw\right)^{\frac{5\varepsilon}{2}} \\ &\hookrightarrow \left(S_{R-\varepsilon}^{R+\varepsilon}(p) \cap P\right)^{\frac{7\varepsilon}{2}} \\ &\hookrightarrow \left(S_{R-\varepsilon}^{R+\varepsilon}(p) \cap \triangle uvw\right)^{\frac{9\varepsilon}{2}} . \end{aligned}$$

By the bounds in Assumption 2 on

- the angle betwen edges at a common vertex,
- the distance between vertices,
- the angles between triangles with a common vertex or edge,
- the distance between vertices and cells they do not intersect with,

the weak feature size of $S_{R-\varepsilon}^{R+\varepsilon}(p) \cap C_p^{\frac{3\varepsilon}{2}}$ is greater than 5ε, and we have the following isomorphism on homology induced by the inclusions above

$$H_\bullet \left(S_{R-\varepsilon}^{R+\varepsilon}(p) \cap |X|\right) \cong H_\bullet \left(\left(S_{R-\varepsilon}^{R+\varepsilon}(p) \cap |X|\right)^{\frac{5\varepsilon}{2}}\right) \cong H_\bullet \left(\left(S_{R-\varepsilon}^{R+\varepsilon}(p) \cap |X|\right)^{\frac{9\varepsilon}{2}}\right).$$

The above homology factors through $\left(S_{R-\varepsilon}^{R+\varepsilon}(p) \cap P\right)^{\frac{3\varepsilon}{2}}$ and $\left(S_{R-\varepsilon}^{R+\varepsilon}(p) \cap P\right)^{\frac{7\varepsilon}{2}}$ so we have

$$\mathrm{rk}_\bullet^{\frac{3\varepsilon}{2}, \frac{7\varepsilon}{2}} \left(S_{R-\varepsilon}^{R+\varepsilon}(p) \cap P\right) = \left|H_\bullet \left(S_{R-\varepsilon}^{R+\varepsilon}(p) \cap |X|\right)\right| .$$

As $\|p - v\| \leq \frac{R}{2} - 2\varepsilon$, we have

$$\left|H_\bullet\left(S^{R+\varepsilon}_{R-\varepsilon}(p) \cap |X|\right)\right| = \left|H_\bullet\left(\partial B_R(v) \cap |X|\right)\right|,$$

giving

$$\mathrm{rk}_\bullet^{\frac{3\varepsilon}{2}, \frac{7\varepsilon}{2}}\left(S^{R+\varepsilon}_{R-\varepsilon}(p) \cap P\right) = \left|H_\bullet\left(\partial B_R(v) \cap |X|\right)\right|.$$

Case 1: $\left|H_0\left(\partial B_R(v) \cap |X|\right)\right| = n$, $\left|H_1\left(\partial B_R(v) \cap |X|\right)\right| = 0$, $n \neq 2$

By the above, we have $\mathrm{Sig}(p) = (n, 0)$, $n \neq 2$, and so the (ε, R)-local structure of P at p is not maximal.

Case 2: $\left|H_0\left(\partial B_R(v) \cap |X|\right)\right| = 2$, $\left|H_1\left(\partial B_R(v) \cap |X|\right)\right| = 0$

By the above, we have $\mathrm{Sig}(p) = (2, 0)$. Let $C^{2\varepsilon}_p$ be the connected component of $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}\left(S^{R+\varepsilon}_{R-\varepsilon}(p) \cap P\right)$ containing p.

Assume that v is a face of some triangle $\triangle uvw$. Then by the bounds placed on angles between edges, and distances between edges without a common face, edges and vertices which are not faces, and vertices and triangles they are not a face of (see Assumption 2), and Lemma 2.2.6 at least one connected component in $\left(S^{R+\varepsilon}_{R-\varepsilon}(p) \cap C^{\frac{3\varepsilon}{2}}_p\right)^{\frac{3\varepsilon}{2}}$ has a diameter greater than $2\sqrt{2}\varepsilon$. Thus, the (ε, R)-local structure of P at p is not maximal.

If v is only the face of edges, then by the bounds placed on angles between edges, and distances between edges without a common face, edges and vertices which are not faces, and vertices and triangles they are not a face of (see Assumption 2), both connected components come from two edges uv and wv, Lemmas 1.1.2, 1.1.3 and 2.2.4 give that the (ε, R)-local structure of P at p is not maximal.

Case 3: $\left|H_0\left(\partial B_R(v) \cap |X|\right)\right| = 1$ $\left|H_1\left(\partial B_R(v) \cap |X|\right)\right| = 1$

Again, we have $\mathrm{Sig}(p) = (1, 1)$ so there are at least three triangles having v as a common vertex. Let p_X be the closest point in $|X|$ to p, and let $x_1 \in \partial B_R(p) \cap |X|$ be colinear with v and p_X, then there is $q_1 \in S^{R+\varepsilon}_{R-\varepsilon} \cap P$ with $\|q_1 - x_1\| \leq \varepsilon$.

Now take any $q_2 \in S_{R-\varepsilon}^{R+\varepsilon} \cap P$, and let x_2 be the point in $|X| \cap \partial B_R(p)$ closest to q_2. Then from Lemma 2.2.1

$$\|q_2 - x_2\| \leq \sqrt{2}\varepsilon.$$

Consider the rays L_1, L_2 from v through x_1, x_2 respectively, and assume $d(p, L_1) \leq \varepsilon$.

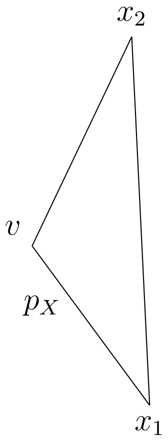


FIGURE 2.9. $d(q_2, H_2) \leq \varepsilon$

We have

$$\|x_1 - v\| = \|x_1 - p_X\| + \|p_X - v\| \leq \frac{3R}{2} - 2\varepsilon,$$
$$\|x_2 - v\| \leq R + \varepsilon,$$

and so

$$\|x_2 - x_1\| = \|x_2 - v\|^2 + \|x_1 - v\|^2 - 2\|x_2 - v\|\|x_1 - v\|^2 \cos \angle x_1 v x_2$$
$$\leq \left(\frac{3R}{2} - 2\varepsilon\right)^2 + (R+\varepsilon)^2 - \left(\frac{3R}{2} - 2\varepsilon\right)(R+\varepsilon) \cos x_1 v x_2.$$

By condition 13 in Assumption 2 the angle between them is bounded above by $\Psi_3(\varepsilon, R)$, so

$$\|x_2 - x_1\| \leq 2\sqrt{R^2 - \varepsilon^2} - (1 + \sqrt{2})\varepsilon,$$

and so

$$\|q_2 - q_1\| \leq 2\sqrt{R^2 - \varepsilon^2} - (2 + 2\sqrt{2})\varepsilon.$$

Thus, the (ε, R)-local structure of P at p is not maximal.

Case 4: $|H_0\left(\partial B_R(v) \cap |X|\right)| = 1$, $|H_1\left(\partial B_R(v) \cap |X|\right)| = n$, $n \geq 2$

By the argument at the start of this proof, $\mathrm{Sig}(p) = (1, n)$, $n \geq 2$ and so the (ε, R)-local structure of P at p is not maximal. □

Next, we bound the region near edges that are not locally maximal.

Proposition 2.3.15. *Let $\overline{uv}$ be an edge of $|X| \subset \mathbb{R}^n$, which is not locally maximal, and let P be an ε-sample of $|X|$. Then, for all $p \in P$ with $d(\overline{uv}, p) \leq \frac{R}{2} - 2\varepsilon$, the (ε, R)-local structure of P at p is not maximal.*

Proof. If an edge $\overline{uv}$ is not locally maximal, then there is at least one triangle $\triangle uvw$.

We consider 3 cases:

1. there is a unique triangle $\triangle uvw$ with $\overline{uv}$ in the boundary,
2. there are exactly two triangles $\triangle uvw_1$ and $\triangle uvw_2$ with $\overline{uv}$ in their boundaries,
3. there are three or more triangles $\triangle uvw_1$, $\triangle uvw_2$ and $\triangle uvw_3$ with $\overline{uv}$ in their boundaries.

Recall that we restrict our attention to the connected components $C_p, C_p^{\frac{3\varepsilon}{2}}$ of $S_{R-\varepsilon}^{R+\varepsilon}(p) \cap |X|$ and $\left(S_{R-\varepsilon}^{R+\varepsilon}(p) \cap P\right)^{\frac{3\varepsilon}{2}}$ which contains p.

By the bounds in Assumption 2 on

- the angle betwen edges at a common vertex,
- the distance between edges that do not have a common face,
- the angles between triangles with a common edge,
- the distance between edges and cells they do not intersect with,

the weak feature size of C_p is greater than 5ε. Hence by the same argument as at the start of the poof of Proposition 2.3.14,

$$\mathrm{Sig}_{\varepsilon,R}(p) = \left(\left|H_0\left(\partial B_R(m) \cap |X|\right)\right|, \left|H_1\left(\partial B_R(m) \cap |X|\right)\right|\right).$$

Thus, in cases 1 and 3, we get $\mathrm{Sig}(p) = (1, 0)$ and $\mathrm{Sig}(p) = (1, n)$ for $n \geq 3$ respectively.

In case 2, we get $\mathrm{Sig}(p) = (1, 1)$, and so need to check the geometric condition. By Lemma 2.2.3, there is a $q_1 \in S^{R+\varepsilon}_{R-\varepsilon}(p) \cap P$ such that for all $q_2 \in S^{R+\varepsilon}_{R-\varepsilon}(p) \cap P$

$$\|q_2 - q_1\| < 2\sqrt{R^2 - \varepsilon^2} - (1 + \sqrt{2})\varepsilon,$$

and so the (ε, R)-local structure of P at p is not maximal.

Hence, in all 3 cases, the (ε, R) local structure of P at p is not maximal.

□

2.4. 2-Complex Algorithm and Correctness

In this section, we present a set of algorithms, which together, recover the structure of X from an ε-sample P of an embedding $(X, \Theta) \subset \mathbb{R}^n$. Theorem 2.4.26 states that given an ε-sample P of an embedded 2 complex $|X|= (X, \Theta_X) \subset \mathbb{R}^n$ satisfying Assumption 2, we can recover the structure of X using this algorithm. There is a sequence of lemmas (Lemmas 2.4.10 to 2.4.25), which culminates in the 'big theorem' (Theorem 2.4.26).

The algorithm partitions P into P_{LM} and P_{NLM}, such that for each $p \in P_{LM}$ the (ε, R)-local structure of P at p is maximal, and for each $p \in P_{NLM}$ the (ε, R)-local structure of P at p is not maximal. We then detect the number of vertices, the number of edges, the number of triangles and the incidence operator. To obtain P_{LM} and P_{NLM}, we use

$$\Delta_{\varepsilon,R} : P \to \{0, 1\},$$

see Algorithm 4.

Let $\mathcal{C}_p$ be the samples $q \in P$ in the connected component containing p in the threshold graph

$$\mathcal{G}_p = \mathfrak{G}_{3\varepsilon}\left(B_{R+\varepsilon}(p) \cap P\right)$$

with $\|q - p\| \in [R - \varepsilon, R + \varepsilon]$. In the definitions of (ε, R)-local structure (Definitions 2.3.4 and 2.3.5), we used

$$\mathrm{rk}_{\bullet}^{\frac{3\varepsilon}{2}, \frac{7\varepsilon}{2}}\left(S^{R+\varepsilon}_{R-\varepsilon}(p) \cap P\right),$$

which by the Nerve Lemma (Corollary 4G.3 [16]) is equal to the rank, $\mathcal{RK}_\bullet$, of the map

$$H_\bullet\left(\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}\left(S_{R-\varepsilon}^{R+\varepsilon}(p)\cap\mathcal{C}_p\right)\right)\to H_\bullet\left(\check{\mathcal{C}}_{\frac{7\varepsilon}{2}}\left(S_{R-\varepsilon}^{R+\varepsilon}(p)\cap\mathcal{C}_p\right)\right)$$

induced by the inclusion

$$\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}\left(S_{R-\varepsilon}^{R+\varepsilon}(p)\cap P\right)\hookrightarrow\check{\mathcal{C}}_{\frac{7\varepsilon}{2}}\left(S_{R-\varepsilon}^{R+\varepsilon}(p)\cap P\right).$$

Hence, $\Delta_{\varepsilon,R}(p)$ returns 0 if the (ε,R)-local structure of P at P is not maximal, and returns 1 if it is maximal. Then,

$$P_{NLM}=\Delta_{\varepsilon,R}^{-1}(0)$$

and

$$P_{NLM}=\Delta_{\varepsilon,R}^{-1}(1).$$

Remark 2.4.1. We can appeal to the Nerve Lemma, as the balls used in the construction of $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}\left(S_{R-\varepsilon}^{R+\varepsilon}(p)\cap\mathcal{C}_p\right)$ and $\check{\mathcal{C}}_{\frac{7\varepsilon}{2}}\left(S_{R-\varepsilon}^{R+\varepsilon}(p)\cap\mathcal{C}_p\right)$ lead us to *good covers* of $\left(S_{R-\varepsilon}^{R+\varepsilon}(p)\cap P\right)^{\frac{3\varepsilon}{2}}$ and $\left(S_{R-\varepsilon}^{R+\varepsilon}(p)\cap P\right)^{\frac{7\varepsilon}{2}}$ respectively. To see that these covers satisfy the 'every non-empty intersection is contractible' condition required to be a good cover, note that we are using the Čech complex, rather than the Viertoris-Rips complex. Combining this with the linearity of the embedding and the assumptions placed on both ε and R, we have covers that satisfy the Nerve Lemma.

After we have P_{LM}, we use the function

$$\mathfrak{D}_{\varepsilon,R}(p):P_{LM}\to\{0,\,1,\,2\},$$

see Algorithm 5 to determine what dimension of (ε,R)-local structure each sample in P_{LM} has.

Recall that our end goal is to learn the combinatorial structure of X. We begin by learning the number of triangles, locally maximal edges, and locally maximal vertices. Consider the following three subsets of P_{LM}:

$$\begin{aligned}P_{LM,2}&=\{p\in P_{LM}\mid\mathfrak{D}_{\varepsilon,R}(p)=2\},\\P_{LM,1}&=\{p\in P_{LM}\mid\mathfrak{D}_{\varepsilon,R}(p)=1\},\\P_{LM,0}&=\{p\in P_{LM}\mid\mathfrak{D}_{\varepsilon,R}(p)=0\}.\end{aligned}$$

Algorithm 4: $\Delta_{\varepsilon,R}(p)$

```
Data: An $\varepsilon$-dense sample $P$ of an embedded 2-complex $|X|$, a
      point $p \in P$.
Result: 0 if the $(\varepsilon, R)$-local structure of $P$ at $p$ is not maximal,
        1 if the $(\varepsilon, R)$-local structure of $P$ at $p$ is maximal.
begin
    $\mathcal{G}_p \longleftarrow \{q \in P \mid \|p - q\| \leq R + \varepsilon\}$;
    connect $q, q' \in \mathcal{G}_p$ if $\|q - q'\| \leq 3\varepsilon$;
    $\mathcal{C}_p \longleftarrow \{q \in \mathcal{G}_p \mid q$ is path connected to $p$ in $\mathcal{G}_p\}$;
    remove $q \in \mathcal{C}_p$ if $\|p - q\| \geq R - \varepsilon$;
    if $\mathcal{RK}_0 = 0$ and $\mathcal{RK}_1 = 0$ then
        return 1
    else if $\mathcal{RK}_0 = 1$ and $\mathcal{RK}_1 \neq 1$ then
        return 0
    else if $\mathcal{RK}_0 = 1$ and $\mathcal{RK}_1 = 1$ then
        if $\forall q_1, q_2 \in C_p$, $\exists q_0$ such that
         $\|q_1 - q_0\|, \|q_2 - q_0\|, \|q_2 - q_1\| \in [\sqrt{3(R^2 - \varepsilon^2)}, \sqrt{3}R]$
         then
            return 1
        else
            return 0
    else if $\mathcal{RK}_0 = 2$ and $\mathcal{RK}_1 = 0$ then
        if $\max\{\mathcal{D}(c_1), \mathcal{D}(c_2)\} \leq 5\varepsilon$ then
            if $\langle q_1 - p, q_2 - p\rangle > -R^2 + 2R\varepsilon - 7\varepsilon^2$ then
                return 1
            else
                return 0
        else
            return 0
    else if $\mathcal{RK}_0 = n,\ n \neq 0, 1, 2$ and $\mathcal{RK}_1 = 0$ then
        return 0
```

When partitioning P into P_{LM} and P_{NLM}, there is a *grey* region where a sample p could be in either of these two sets. This presents a problem for learning the combinatorics of X from the partitioning P_{LM} and P_{NLM}. We can overcome this, by *cleaning* P_{LM}. In particular, we clean $P_{LM,2}$ and $P_{LM,1}$.

Algorithm 5: $\mathfrak{D}_{\varepsilon,R}(p)$

```
Data: An $\varepsilon$-dense sample $P$ of an embedded 2-complex $|X|$, a
      point $p \in P$ such that the $(\varepsilon, R)$-local structure of $P$ at $p$ is
      maximal.
Result: 0 if the $(\varepsilon, R)$-local structure of $P$ at $p$ is maximal of
        dimension 0,
        1 if the $(\varepsilon, R)$-local structure of $P$ at $p$ is maximal of
        dimension 1,
        2 if the $(\varepsilon, R)$-local structure of $P$ at $p$ is maximal of
        dimension 2.
begin
    $\mathcal{G}_p \longleftarrow \{q \in P \mid \|p - q\| \leq R + \varepsilon\}$;
    connect $q, q' \in \mathcal{G}_p$ if $\|q - q'\| \leq 3\varepsilon$;
    $\mathcal{C}_p \longleftarrow \{q \in \mathcal{G}_p \mid q$ is path connected to $p$ in $\mathcal{G}_p\}$;
    remove $q \in C_p$ if $\|p - q\| \leq R - \varepsilon$;
    if $\mathcal{RK}_0 = 0$ and $\mathcal{RK}_1 = 0$ then
        return 0
    else if $\mathcal{RK}_0 = 2,\ n \neq 0, 1, 2$ and $\mathcal{RK}_1 = 0$ then
        return 1
    else if $\mathcal{RK}_0 = 1,\ n \neq 0, 1, 2$ and $\mathcal{RK}_1 = 1$ then
        return 2
```

We begin by introducing the notion of a connected component of $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,1})$ *spanning* an edge, and then introduce the notion of a connected component of $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,2})$ *spanning* a triangle.

Definition 2.4.2 (Spanning an edge)**.** *We say a connected component of* $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,1})$ spans *a locally maximal edge* $\overline{uv}$ *if it contains a sample* p *within* ε *of the midpoint of* $\overline{uv}$*.*

Definition 2.4.3 (Spanning a triangle)**.** *We say a connected component of* $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,2})$ spans *a triangle* $\triangle uvw$ *if it contains a sample* p *within* ε *of the midpoint of* $\triangle uvw$*.*

We require some geometric conditions on when a connected component spans an edge or a triangle. For an edge, we will use the diameter of the connected component as a condition.

Proposition 2.4.4. *A connected component* C *of* $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,1})$ spans *a locally maximal edge* $\overline{uv}$ *if and only if* $\mathcal{D}(C) \geq \frac{3R}{2} - 2\varepsilon$*.*

Proof. Let C be a connected component of $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,1})$ which spans a locally maximal edge $\overline{uv}$, with midpoint m_{uv}. Then, there is a sample $p_m \in C$ such that $\|p_m - m_{uv}\| \leq \varepsilon$.

To show that $\mathcal{D}(C) \geq \frac{9R}{2}$, we show that there are two points $x_u, x_y \in \overline{uv}$ such that

1. $\|u - x_u\| > \frac{3R}{2} + 2\varepsilon$,
2. $\|v - x_v\| > \frac{3R}{2} + 2\varepsilon$,
3. $\|x_u - x_v\| \geq \frac{3R}{2}$.

Without loss of generality, we show that x_u exists, and

$$\|x_u - m_{uv}\| \geq \frac{3R}{4} + \varepsilon.$$

By Assumption 2, $\|u - v\| \geq 6(R + \varepsilon)$. As $\overline{uv}$ is a line segment, for all $\eta \in [0, \frac{9R}{4} + 3\varepsilon]$ there is a point $x_\eta \in \overline{uv}$ such that $\|x_\eta - u\| = \eta$. Letting $\eta = \frac{3R}{2} + 2\varepsilon$, there is a point, namely x_u such that $\|x_u - u\| = \frac{3R}{2} + 2\varepsilon$. As P is an ε-sample, there is a sample p_u such that $\|x_u - p_u\| \leq \varepsilon$, and hence $\|p_u - u\| > \frac{3R}{2} + \varepsilon$. Thus, the (ε, R)-local structure of P at p_u is maximal of dimension 1.

We can repeat this argument for all $\eta \in [\frac{3R}{2} + 2\varepsilon, \frac{9R}{4} + 3\varepsilon]$, and obtain a path of points $x_\eta \in \overline{uv}$ and samples $p_\eta \in P$ connecting p_u to p_m.

This also holds when we replace u with v, and hence we have p_u and p_v. Finally, we have

$$\begin{aligned}\|p_u - p_v\| \geq & \|x_u - x_v\| - \|p_u - x_u\| - \|p_v - x_v\| \\ \geq & \frac{3R}{2} - 2\varepsilon,\end{aligned}$$

and hence $\mathcal{D}(C) \geq \frac{3R}{2} - 2\varepsilon$.

Now, we show that if $\mathcal{D}(C) \geq \frac{3R}{2} - 2\varepsilon$, then C spans some locally maximal edge.

If $\mathcal{D}(C) \geq \frac{3R}{2} - 2\varepsilon$, then there are points $p, q \in C$ with

$$\|p - q\| \geq \frac{3R}{2} - 2\varepsilon.$$

As P is an ε-sample of $|X|$, there are points $x_p, x_q \in |X|$, with

$$\|x_p - p\|,\ \|x_q - q\| \leq \varepsilon.$$

Let m_{pq} be the midpoint of x_p and x_q. As p and q are in the same connected component of $\check{C}_{\frac{3\varepsilon}{2}}\left(P_{LM,1}\right)$, we know there is a sequence of points $\{q_i\}_{i=0}^{m}$ with $q_0 = p,\ q_m = q$ and for all $0 < i \leq m$, $\|q_i - q_{i-1}\| \leq 3\varepsilon$. Again, P is an ε-sample of $|X|$, and as $q_i \in P_{LM,1},\ \forall 0 \leq i \leq m$, for each q_i there is some $x_i \in |X|$ which is on a locally maximal edge, and $\|q_i - x_i\| \leq \varepsilon$. From Assumption 2 and Proposition 2.3.11, there is a locally maximal edge, say $\overline{uv}$, such $x_i \in \overline{uv}, \forall 0 \leq i \leq m$. Let the midpoint of $\overline{uv}$ be x_{uv}.

We now split into two cases:

I there is some i such that $x_i = x_{uv}$,
II for all i we have $x_i \neq x_{uv}$.

Case I: The connected component C is a spanning connected component, as it contains a sample which is within ε of the midpoint x_{uv} of the locally maximal edge $\overline{uv}$.

Case II: As no q_i is within ε of m_{uv}, we know that $q_i\ \forall 0 \leq i \leq m$ are on the same side of $\overline{uv}$. That is, for all q_i, without loss of generality,

$$\|q_i - x_{uv}\| \leq \|q_i - u\| \geq \frac{3\sqrt{3}}{2}R + 3\varepsilon$$
$$\|q_m - v\| \geq \frac{3R}{2} + \varepsilon.$$

Further, assume that

$$\|q_0 - m_{uv}\| \leq \|q_m - x_{uv}\|.$$

There is another sequence of points $\{x'_j\}_{j=0}^{m'}$ in $\overline{uv}$ with $x'_0 = x_m$ and $x'_{m'} = x_{uv}$, and for $0 < j \leq m'$

$$\|x'_j - x'_{j-1}\| \leq \varepsilon.$$

Then, there exists $q'_j \in P$ with

$$\|q'_j - x'_j\| \leq \varepsilon$$
$$\|q'j - q'_{j-1}\| \leq \varepsilon \, \forall 0 < j \leq m'$$
$$\|q'_j - v\| \geq \frac{3R}{2} + \varepsilon.$$

By Assumption 2 and Proposition 2.3.11, $q'_j \in P_{LM,1}$ for all $0 \leq j \leq m'$. Hence, each q'_j is in the same connected component C as q_m.

Thus, C contains a sample $q'_{m'}$ which is within ε of the midpoint of the locally maximal edge $\overline{uv}$. Hence, C is a spanning connected component.

Thus a component C of $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,1})$ spans a locally maximal edge $\overline{uv}$ if and only if $\mathcal{D}(C) \geq \frac{3R}{2} - 2\varepsilon$. □

Unfortunately, it is not immediately clear that such a test is suitable for detecting components that span triangles. For instance, consider a complex that consists of a single triangle, its three edges, and the three required vertices. While heuristically, it is unlikely to occur, the sampling could lead to 2 connected components $C_1, C_2 \in \check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,2})$: one which is far away from the boundary of the triangle, and one that is *surrounded* by points in P_{NLM}, both with large diameters. In fact, the one we wish to say is spanning, say C_1, will have a smaller diameter than the other one, C_2. Note, however, that as C_2 does not contain a sample p near the midpoint of $\triangle uvw$, if $\mathcal{D}(C_1) \leq \mathcal{D}(C_2)$, then C_2 contains a non-contractible loop. However, a sample $p \in P$ near the midpoint $m_{\triangle uvw}$ of a triangle $\triangle uvw$ is not near any samples $q \notin P_{LM,2}$, and so we can exploit this fact to obtain a geometric test.

Proposition 2.4.5. *A connected component C of $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}$* spans *a triangle $\triangle uvw$ if and only if there is a point $p \in C$ such that*

$$B_{\frac{R}{2}+\varepsilon}(p) \cap P \subset P_{LM,2}.$$

Proof. First, let C be a connected component of $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}$ which spans some triangle $\triangle uvw$ with midpoint m. As P is an ε-sample of X, there is a sample $p_m \in P$ with $\|p_m - m\| \leq \varepsilon$. As the radius of the inscribed circle of $\triangle uvw$ is at least $2R + 3\varepsilon$, m is at least $2R + 3\varepsilon$ from $\partial \triangle uvw$. Thus, $d(p_m, \partial \triangle uvw) \geq 2R + 2\varepsilon$.

Hence, for all $q \in B_{\frac{R}{2}+2\varepsilon}(p) \cap P$, $d(q, \partial\triangle uvw) \geq \frac{3R}{2} + \varepsilon$, and so $q \in P_{LM,2}$.

Now, take $p \in P_{LM,2}$ such that $B_{\frac{R}{2}+\varepsilon}(p) \cap P \subset P_{LM,2}$. Then, there is some triangle $\triangle uvw$ with $d(\triangle uvw, p) \leq \varepsilon$. As $p \in P_{LM,2}$, we know that $d(\partial\triangle uvw, p) > \frac{R}{2} - \varepsilon$. By assumption, for all $q \in B_{\frac{R}{2}+\varepsilon}(p) \cap P$, we have $d(\partial\triangle uvw, q) > \frac{R}{2} - \varepsilon$. Recall that P is an ε-sample of $|X|$, so there is a point $x \in X$ such that $\|p - x\| \leq \varepsilon$. As $\triangle uvw$ is convex, and every $B_{\frac{R}{2}+\varepsilon}(p) \cap P \subset P_{LM,2}$, we have

$$d(\partial\triangle uvw, x) \geq \frac{R}{2} + 2\varepsilon + \frac{R}{2} - 2\varepsilon = R.$$

Hence, is a point $y \in B_{\frac{R}{2}+2\varepsilon}(p) \cap \triangle uvw$ with

$$d(\partial\triangle uvw, y) \geq \frac{R}{2} + 2\varepsilon,$$

and a sample $q \in B_{\frac{R}{2}+2\varepsilon}(p) \cap P_{LM,2}$ with $\|q - y\| \leq \varepsilon$.

Now, we can construct a sequence of points $\{y_i\}_{i=0}^{m} \subset \triangle uvw$ such that $\|y_i - y_{i-1}\| \leq \varepsilon$ for $1 \leq i \leq m$, and $y_0 = x$, $y_m = y$. Further, for each y_i there is a $q_i \in P$ with $\|q_i - y_i\| \leq \varepsilon$, and $q_i \in P_{LM,2}$. Note, that this means p and q_m are in the same connected component C of $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,2})$.

Finally, we construct a similar sequence of points $\{\widetilde{y_j}\}_{j=0}^{\widetilde{m}}$ in $|X|$ from y to $m_{\triangle uvw}$ with $\widetilde{y_0} = y$, $\widetilde{y_{\widetilde{m}}} = m_{\triangle uvw}$. Again, for each $\widetilde{y_j}$, there is a $\widetilde{q_j} \in P$ with $\|\widetilde{y_j} - \widetilde{q_j}\| \leq \varepsilon$ and $\widetilde{q_j} \in P_{LM,2}$. Hence, the $\widetilde{q_j}$ are in the same connected component of $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,2})$, and further, this connected component is C. □

We now have geometric conditions for determining if a connected component of $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,2})$/$\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,1})$ spans a triangle/edge respectively. Next, show that the locally maximal vertices of X are in bijection with connected components of $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,0})$, the locally maximal edges of X are in bijection with the spanning connected components of $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,1})$, and that the triangles of X are in bijection with the spanning connected components of $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,2})$.

We begin with the locally maximal vertices.

Proposition 2.4.6. *The connected components of $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,0})$ are in bijection with the set V_{LM} of locally maximal vertices of X.*

Proof. Let V_{LM} be the set of locally maximal vertices of X. Let v be a locally maximal vertex, then by Proposition 2.3.10, $\forall p \in P$ with $\|p - v\| \leq 4\varepsilon$, $p \in P_{LM,0}$. In fact, by Assumption 2, any $p \in P$ with $\|p - v\| \leq 4\varepsilon$ is actually within ε of v. Hence, every $p \in P_{LM,0}$ within ε of v are in the same connected component of $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,0})$.

Now, take a connected component C of $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,0})$. Each $p \in C$ is within ε of a locally maximal vertex v_p of X. By Assumption 2, every locally maximal vertex v is at least 5ε away from any other cell of X, and hence $\forall p \in C$, v_p is the same.

Hence, the connected components of $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,0})$ correspond bijectively to the locally maximal vertices of X. □

Next, we show that the edge-spanning components are in bijection with the locally maximal edges.

Proposition 2.4.7. *The* spanning *components of* $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,1})$ *are in bijection with the set* E_{LM} *of locally maximal edges of* X.

Proof. Let $E_{LM} \subset E$ be the set of locally maximal edges in X. By Proposition 2.4.4, a connected component C of $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,1})$ spans an edge $\overline{uv}$ if and only if it contains a sample p within ε of the midpoint m of $\overline{uv}$.

If a connected component C is a spanning component, then there is some locally maximal edge $\overline{uv}$ with midpoint m such that there is a sample $p \in C$ with $\|m - p\| \leq \varepsilon$.

For any locally maximal $\overline{uv} \in E_{LM}$ with midpoint m, there is some sample $p \in P$ such that $\|m - p\| \leq \varepsilon$. Then, by Assumption 2 and proposition 2.3.11, $p \in P_{LM,1}$, and so there is some spanning connected component $C_{\overline{uv}}$ in $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,1})$.

Now, consider a locally maximal edge $\overline{uv'}$, $v' \neq v$, and take samples $p, q \in P_{LM,2}$ such that $d(\overline{uv}, p),\ d(\overline{uv'}, q) \leq \varepsilon$. By Assumption 2, $\|p - q\| > 6\varepsilon$, and so p and q are in different connected components of $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,1})$.

Finally, consider a locally maximal edge $\overline{u'v'}$ such that $\overline{uv}$ and $\overline{u'v'}$ do not have a common vertex. Take samples $p, q \in P_{LM,2}$ such that

$$d(\overline{uv}, p),\ d(\overline{u'v'}, q) \leq \varepsilon.$$

Again, by Assumption 2, $\|p - q\| > 6\varepsilon$, and so p and q are in different connected components of $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,1})$.

Hence, each connected component C only consists of samples p with $d(\overline{uv}, p) \leq \varepsilon$ for a single locally maximal edge $\overline{uv}$.

Thus, the spanning connected components of $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,2})$ are in bijection with the locally maximal edges of $|X|$. □

Finally, we show that the spanning components of $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,2})$ are in bijection with the triangles of X.

Proposition 2.4.8. *The* spanning *components of* $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,2})$ *are in bijection with the set* T *of triangles in* X.

Proof. From Proposition 2.4.5, a connected component C of $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,2})$ spans a triangle $\triangle uvw$ if and only if it contains a sample p within ε of the midpoint m of $\triangle uvw$.

As P is a ε-sample of $|X|$, for every $\triangle uvw$ with midpoint m, there is a sample $p \in P$ such that $\|p - m\| \leq \varepsilon$. Hence, there is a spanning connected component C in $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,2})$.

Now, consider C a spanning component of $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,2})$. Then, as P is a ε-sample, there is some $\triangle uvw$ with midpoint m such that there is a sample $p \in C$ with $\|p - m\| \leq \varepsilon$.

Consider two triangles $\triangle uvw$, $\triangle u'v'w'$, and take two samples $p, p' \in P_{LM,2}$ with

$$d(\triangle uvw, p),\ d(\triangle u'v'w', p') \leq \varepsilon.$$

As $p, p' \in P_{LM,2}$, we know that

$$d(\partial\triangle uvw, p),\ d(\partial\triangle u'v'w', p') > R + \varepsilon,$$

and so by Assumption 2, $\|p - p'\| > 6\varepsilon$.

Hence, the spanning components of $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,2})$ are in bijection with the triangles of X. □

Having identified the locally maximal cells X_{LM} of X, we could learn the combinatorial structure of X by identifying the structure of X_{NLM} from P_{NLM}, and combining this with what we know about X_{LM} from P_{LM}. The process in Chapter 1 could be applied, but this requires the existence of some $\widetilde{\varepsilon}$ such that P_{NLM} is a $\widetilde{\varepsilon}$-sample of X_{NLM} satisfying Assumption 1. This would impose stricter assumptions than Assumption 2, but after ensuring these new assumptions are satisfied, works out of the box.

To avoid placing stricter assumptions on $|X|$, we use the idea of *witness points* to discover the combinatorics. For each sample $p \in P_{NLM}$, we can examine the spanning connected components C_{LM} of $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,1})$ and $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,2})$ such that $C_{LM} \cap B_{R+3\varepsilon}(p) \neq \emptyset$. In particular, we can use $\mathfrak{D}_{\varepsilon,R}(q)$ for some $q \in C_{LM}$, to determine of what dimension the local structure is maximal. If there is a q in $C_{LM} \cap S_{R-\varepsilon}^{R+\varepsilon}(p)$ such that $\mathfrak{D}_{\varepsilon,R}(q) = 1$, then p is near a vertex.

If there are no connected components C_{LM} which are (ε, R)-locally maximal of dimension 1, then p only *witnesses* samples $q \in P_{LM}$ such that the (ε, R)-local structure of P at q is maximal of dimension 2. Hence, we need to understand the combinatorics of $|X| \setminus (E_{LM} \cup V_{LM})$ where E_{LM} is the set of locally maximal edges and V_{LM} the set of locally maximal vertices.

In Assumption 2, we assumed that for any triangle $\triangle uvw$,

$$\angle uvw, \angle vwu, \angle wuv \geq \frac{\pi}{6}.$$

This means that for any sample $p \in P_{NLM}$ with $d(\partial\triangle uvw, p) < R + \varepsilon$ for some $\triangle uvw$, there is some sample $q \in P_{LM,2}$ with $d(\triangle uvw, q) \leq \varepsilon$ and $d(\partial\triangle uvw, p) \geq R + \varepsilon$, such that $\|q - p\| \leq \frac{2\sqrt{2}(R+2\varepsilon)}{\sqrt{3}-1}$. Further, q is in a triangle spanning component $\mathcal{T}$.

Similarly, for any sample $p \in P_{NLM}$ with $d(\partial\overline{uv}, p) < \frac{3R}{2} + \varepsilon$ for some edge $\overline{uv}$, there is a sample $q \in P_{LM,1}$ with $d(\partial\overline{uv}, p) \geq \frac{3R}{2} + \varepsilon$ such that $\|q - p\| \leq \frac{2\sqrt{2}(R+2\varepsilon)}{\sqrt{3}-1}$. Further, q is in an edge spanning component $\mathcal{E}$.

This leads us to say a sample $p \in P_{NLM}$ *witnesses* a spanning connected component $\mathcal{C}$ if

$$B_{\frac{2\sqrt{2}(R+2\varepsilon)}{\sqrt{3}-1}}(p) \cap \mathcal{C} \neq \emptyset.$$

For ease of reading, we set $\kappa = \frac{2\sqrt{2}}{\sqrt{3}-1}$.

Definition 2.4.9 (Witnessing a spanning component)**.** *Let P be an ε-sample P of an embedded 2-complex $|X|$ satisfying Assumption 2. Then a sample*

$p \in P_{NLM}$ witnesses *an edge/triangle spanning component* $\mathcal{C}$ *if*

$$B_{\kappa(R+\varepsilon)}(p) \cap \mathcal{C} \neq \emptyset.$$

To determine the final combinatorial structure of X, we look at the local neighbourhood of each $p \in P_{NLM}$ and look at both

$$B_{(R+2\varepsilon)\kappa}(p) \cap \check{\mathcal{C}}_{\frac{3\varepsilon}{2}}\left(P_{LM,1}\right)$$
$$B_{(R+2\varepsilon)\kappa}(p) \cap \check{\mathcal{C}}_{\frac{3\varepsilon}{2}}\left(P_{LM,2}\right).$$

If

$$B_{(R+2\varepsilon)\kappa}(p) \cap \check{\mathcal{C}}_{\frac{3\varepsilon}{2}}\left(P_{LM,1}\right) \neq \emptyset$$

then we know that p is *near* a vertex, and the spanning components $\mathcal{E}$ of $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}\left(P_{LM,1}\right)$ that p witnesses, share a boundary vertex. Further, if

$$B_{(R+2\varepsilon)\kappa}(p) \cap \check{\mathcal{C}}_{\frac{3\varepsilon}{2}}\left(P_{LM,2}\right) \neq \emptyset$$

as well, then there are spanning components of $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}\left(P_{LM,2}\right)$ that p witnesses, which have a vertex in common with the edges.

If only

$$B_{(R+2\varepsilon)\kappa}(p) \cap \check{\mathcal{C}}_{\frac{3\varepsilon}{2}}\left(P_{LM,2}\right) \neq \emptyset$$

we examine how many spanning components $\mathcal{T}$ are seen by p, as well as if samples $p \in P_{NLM}$ that witness $\mathcal{T}$, also witness any other spanning components $\mathcal{T}'$. We use this information to partition P_{NLM} into $\{P_i\}$ in Algorithm 8, with a final clean of the partitions, to account for some special cases. As $R \leq 16\varepsilon$, for all $p \in P_{NLM}$ there is some spanning connected component $\mathcal{C}$ such that $B_{\frac{R+\varepsilon}{\kappa}}(p) \cap \mathcal{C} \neq \emptyset$.

We then label each component P_i as follows, from Algorithms 10 and 11:

- -1 if P_i corresponds to 2 vertices,
- 0 if P_i corresponds to a vertex,
- 1 if P_i corresponds to a vertex and an edge,
- 2 if P_i corresponds to two vertices and an edge,
- 3 if P_i corresponds to just an edge,

- 4 if P_i corresponds to two edges and a vertex,
- 5 if P_i corresponds to three edges and two vertices,
- 6 if P_i corresponds to three edges and a vertex,
- 7 if P_i corresponds to three edges and three vertices,
- 8 if P_i corresponds to three edges,
- 9 if P_i corresponds to two edges,

```
Algorithm 6: Spanning triangle components
  Data: Parameters $\varepsilon$, $R$ and $P_{LM,1}$.
  Result: The set of triangle spanning components.
  begin
      Initialise empty set $T$;
      Let $C$ be the set of connected components of $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,2})$;
      for $\mathcal{T} \in C$ do
          if $\exists p \in \mathcal{C}$ such that $B_{R/2+\varepsilon}(p) \cap P \subset P_{LM,2}$ then
              Add $\mathcal{T}$ to $T$;
      return $T$
```

```
Algorithm 7: Spanning edge components
  Data: Parameters $\varepsilon$, $R$ and $P_{LM,1}$.
  Result: The set of triangle spanning components.
  begin
      Initialise empty set $E$;
      Let $C$ be the set of connected components of $\check{\mathcal{C}}_{\frac{3\varepsilon}{2}}(P_{LM,2})$;
      for $\mathcal{E} \in C$ do
          if $\mathcal{D}(\mathcal{T}) \geq \frac{3R}{2} - 2\varepsilon$ then
              Add $\mathcal{E}$ to $E$;
      return $E$
```

The following Lemma's together show that Algorithms 8, 10 and 11 correctly partition P_{NLM} and label the partitions P_i appropriately.

Lemma 2.4.10. *Let $\overline{uv}$ be a locally maximal edge of X, such that u, v are only faces of $\overline{uv}$. Then, there is a unique partition P_1 of P_{NLM} which witnesses $\mathcal{E}$, where $\mathcal{E}$ is the edge spanning component corresponding to $\overline{uv}$. Further, P_1 is assigned label -1 by Algorithms 10 and 11.*

Algorithm 8: Partitioning P_{NLM}

```
Data: An $\varepsilon$-dense sample $P$ of an embedded 2-complex $|X|$,
      partitioned into $P_{NLM},\ P_{LM,0},\ P_{LM,1},\ P_{LM,2}$.
Result: A partition $\{P_i\}$ of $P_{NLM}$, and for each $P_i$, two sets
        $S_E(P_i),\ S_T(P_i)$.
begin
    For each $p \in P_{NLM}$, find all the edge spanning components $\mathcal{E}$
      such that $\mathcal{E} \cap B_{(R+2\varepsilon)\kappa}(p) \neq \emptyset$, and place them in $S_E(p)$;
    Find all the triangle spanning components $\mathcal{T}$ such that
      $\mathcal{T} \cap B_{(R+2\varepsilon)\kappa}(p) \neq \emptyset$, and place them in $S_T(p)$;
    Partition $P_{NLM}$ into $\{\ P_i\ \}$ such that for each $p, q \in P_i$,
      $S_E(p) = S_E(q)$ and $S_T(p) = S_T(q)$;
    Assign $S_E(P_i)$ and $S_T(P_i)$ to each $P_i$;
    for $P_i$ and $P_j$ with $S_E(P_j) \subseteq S_E(P_i)$ and $S_T(P_j) \subseteq S_T(P_i)$ do
        if $S_E(P_j), S_T(P_j) \neq \emptyset$ then
            Merge $P_j$ into $P_i$ with labels $S_E(P_i),\ S_T(P_i)$;
        else if $|S_T(P_j)| \geq 2$ and $\forall p \in P_j$ such that
          $\mathrm{Sig}_{\varepsilon,R}(p) = (n, 0),\ n \in \mathbb{Z}_{\geq 0}$ then
            Merge $P_j$ into $P_i$ with labels $S_E(P_i),\ S_T(P_i)$;
    return $\{P_i\}$, and $S_E(P_i),\ S_T(P_i)$ for each $P_i$
```

Algorithm 9: Order $\{P_i\}$

```
Data: An $\varepsilon$-dense sample $P$ of an embedded 2-complex $|X|$,
      partition $\{P_i\}$ of $P_{NLM}$ with two sets $S_E(P_i),\ S_T(P_i)$ for
      each $P_i$ and partitions of $P_{LM,0},\ P_{LM,1},\ P_{LM,2}$.
Result: Two sets $P^1, P^2 \subset \{\ P_i\ \}$.
begin
    Initialise empty $P^1$ and $P^2$;
    for $P_i \in \{P_i\}$ do
        if $S_E(P_i) \neq \emptyset$ then
            Add $P_i$ to $P^1$
        else if $\exists p \in P_i$ such that $\mathrm{Sig}(p) \neq (1, n)$ then
            Add $P_i$ to $P^1$
        else if $|S_T(P_i)| \neq 1$ then
            Add $P_i$ to $P^1$
        else
            Add $P_i$ to $P^2$
    return $P^1,\ P^2$
```

```
Algorithm 10: Classification of $P^1$
  Data: An $\varepsilon$-dense sample $P$ of an embedded 2-complex $|X|$, $P^1$,
        and partitions of $P_{NLM}$, $P_{LM,0}$, $P_{LM,1}$, $P_{LM,2}$.
  Result: A labeled list $C$, where the label for $P_i$ is $-1$ if $P_i$
          corresponds to 2 vertices, 0 if $P_i$ corresponds to a vertex, 1
          if $P_i$ corresponds to a vertex and an edge, 2 if $P_i$
          corresponds to two vertices and an edge, 3 if $P_i$
          corresponds to just an edge.
  begin
    Initialise empty list $C$;
    for $P_i \in P^1$ do
      if $|S_E(P_i)|= 1$ and $S_T(P_i) = \emptyset$ then
        if $\mathcal{E} \notin S_E(P_j) \forall P_j \neq P_i$ then
          Add $P_i$ to $C$ with label $-1$;
        else if $\exists P_j \neq P_i$ such that $\mathcal{E} \in S_E(P_j)$ then
          Add $P_i$ to $C$ with label 0;
      else if $S_E(P_i) \neq \emptyset$ then
        Add $P_i$ to $C$ with label 0;
      else
        for $\mathcal{T} \in S_T(P_i)$ do
          Let $LN(\mathcal{T}) = \{P_k \mid \mathcal{T} \in S_T(P_k)\}$
        Let $N(P_i) = \bigcap_{\mathcal{T} \in S_T(P_i)} LN(\mathcal{T})$;
        if $N(P_i) = \{P_i, P_k\}$ then
          Add $P_i$ to $C$ with label $1$;
          Add $P_k$ to $C$ with label $0$, unless $P_k$ is already in $C$;
        else if $N(P_i) = \{P_i, P_k, P_l\}$ then
          Add $P_i$ to $C$ with label $3$;
          Add $P_k$ to $C$ with label $0$, unless $P_k$ is already in $C$;
          Add $P_l$ to $C$ with label $0$, unless $P_l$ is already in $C$;
    if $\exists P_i \in P^1 \setminus C$ then
      Add $P_i$ to $C$ with label 2;
    return $C$
```

Proof. As $\overline{uv}$ is a locally maximal edge, there is a corresponding edge spanning component $\mathcal{E}$. As u, v are not faces of any other cell $\sigma \in X$, by Assumption 2 and Propositions 2.3.11 and 2.3.14, the points $p \in P_{NLM}$ which

Algorithm 11: Classification of P^2

```
Data: An $\varepsilon$-dense sample $P$ of an embedded 2-complex $|X|$, $P^2$,
      and partitions of $P_{NLM}$, $P_{LM,0}$, $P_{LM,1}$, $P_{LM,2}$, a labelled list
      $C$ obtained from Algorithm 10.
Result: A labelled list $C$
begin
    for $P_i \in P^2$ do
        if $P_i \notin C$ then
            Let $LN = \{P_k \mid \mathcal{T} \in S_T(P_k)\}$;
            if $LN \cap P^2 = \{P_i, P_k, P_l\}$ then
                Add $P_i, P_k, P_l$ to $C$ with label 3;
            else if $LN \cap P^2 = \{P_i, P_k\}$ then
                Add $P_i$ to $C$ with label 3;
                Add $P_l$ to $C$ with label 4;
            else if $LN \cap P^2 = \{P_i\}$ then
                if $LN = \{P_i\}$ then
                    Add $P_i$ to $C$ with label 7;
                else if $LN = \{P_i, P_k\}$ and $P_k$ has label 0 then
                    Add $P_i$ to $C$ with label 5;
                else if $LN = \{P_i, P_k\}$ and $P_k$ has label 2 then
                    Add $P_i$ to $C$ with label 4;
                else if $LN = \{P_i, P_k, P_l\}$ and $P_k$ has label 0, $P_l$
                  label 1 then
                    Add $P_i$ to $C$ with label 4;
                else if $LN = \{P_i, P_k, P_l\}$ and $P_k$ has label 1, $P_l$
                  label 2 then
                    Add $P_i$ to $C$ with label 3;
                else if $LN = \{P_i, P_k, P_l\}$ and $P_k$ has label 0, $P_l$
                  label 0 then
                    Add $P_i$ to $C$ with label 6;
                else if $LN = \{P_i, P_k, P_l, P_j\}$ and $P_k, P_l, P_j$ have
                  label 0 then
                    Add $P_i$ to $C$ with label 8;
                else if $LN = \{P_i, P_k, P_l, P_j, P_m\}$ and $P_k, P_l, P_j$
                  have label 0 and $P_m$ has label 3 then
                    Add $P_i$ to $C$ with label 9;
    return $C$
```

Algorithm 12: Number of triangles, edges and vertices.

```
Data: An $\varepsilon$-dense sample $P$ of an embedded 2-complex $|X|$,
      partitions of $P_{NLM}$, $P_{LM,0}$, $P_{LM,1}$, $P_{LM,2}$ and the labelled
      list $C$ from Algorithm 11.
Result: The triangles, edges, and vertices in $X$.
begin
    Initialise an empty weighted graph $B$;
    $\forall$ spanning components $\mathcal{T}$ of $P_{LM,2}$, add weight $2$ node to $B$,
      labelled with $\mathcal{T}$;
    $\forall$ spanning components $\mathcal{E}$ of $P_{LM,1}$, add weight $1$ node to $B$,
      labelled with $\mathcal{E}$;
    $\forall$ components $\mathcal{V}$ of $P_{LM,0}$, add weight $0$ node to $B$, labelled
      with $\mathcal{V}$;
    for $P_i \in C$ do
        if $P_i$ has label $-1$ then
            Add $2$ weight $0$ nodes to $B$, labelled with $P_i$;
        else if $P_i$ has label $0$ then
            Add weight $0$ node to $B$, labelled with $P_i$;
        else if $P_i$ has label $1$ then
            Add $2$ weight $0$ nodes to $B$, labelled with $P_i$;
            Add weight $1$ node to $B$, labelled with $P_i$;
        else if $P_i$ has label $2$ then
            Add weight $0$ node to $B$, labelled with $P_i$;
            Add weight $1$ node to $B$, labelled with $P_i$;
        else if $P_i$ has label $3$ then
            Add two weight $0$ nodes to $B$, labelled with $P_i$;
            Add weight $1$ node to $B$, labelled with $P_i$;
        else if $P_i$ has label $4$ then
            Add weight $1$ node to $B$, labelled with $P_i$;
        else if $P_i$ has label $5$ then
            Add weight $0$ node to $B$, labelled with $P_i$;
            Add two weight $1$ nodes to $B$, labelled with $P_i$;
        else if $P_i$ has label $6$ then
            Add two weight $0$ nodes to $B$, labelled with $P_i$;
            Add three weight $1$ nodes to $B$, labelled with $P_i$;
        else if $P_i$ has label $7$ then
            Add three weight $0$ nodes to $B$, labelled with $P_i$;
            Add three weight $1$ nodes to $B$, labelled with $P_i$;
```

witness $\mathcal{E}$ do not witness any other edge spanning component $\mathcal{E}'$ or any triangle spanning component $\mathcal{T}$.

Thus, there is a single partition P_1 of P_{NLM} which contains all the samples p that witness $\mathcal{E}$. By Assumption 2, there is no other partition P_2 of P_{NLM} that witnesses $\mathcal{E}$. Hence, P_1 is assigned label -1. □

Lemma 2.4.11. *Let $\overline{uv}$ be a locally maximal edge of X, such that u and/or v is the face of some locally maximal cell $\sigma \in X$, $\sigma \neq \overline{uv}$. Then, there are partitions P_1, P_2 of P_{NLM}, which witness $\mathcal{E}$, where $\mathcal{E}$ is the edge spanning component corresponding to $\overline{uv}$. Further, P_1 and P_2 are assigned label 0 by Algorithms 10 and 11.*

Proof. As $\overline{uv}$ is a locally maximal edge, there is a corresponding edge spanning component $\mathcal{E}$. Without loss of generality, assume v is the face of some locally maximal cell $\sigma \neq \overline{uv}$.

By Assumption 2 and Propositions 2.3.11, 2.3.12 and 2.3.14, there are samples $p_u, pv_v \in P_{NLM}$ such that

$$\|p_u - u\|,\ \|p_v - v\| \leq \varepsilon.$$

Further, there is a spanning connected component $\mathcal{C}$ which p_v also witnesses but p_u does not witness. Hence, there are two partitions P_v, P_u which witness $\mathcal{E}$. By Assumption 2 and Algorithm 8, there are no other partitions which witness $\mathcal{E}$.

Hence, both P_v and P_u are labelled with 0 by Algorithms 10 and 11. □

Lemma 2.4.12. *Let $\triangle uvw$ be a triangle of X, such that for all locally maximal cells $\sigma \in X$ with $\sigma \neq \triangle uvw$, we have*

$$u,\ v,\ w \notin \sigma.$$

Then, there is a unique partition P_1 of P_{NLM} which witnesses $\mathcal{T}$, where $\mathcal{T}$ is the edge spanning component corresponding to $\triangle uvw$. Further, P_1 is given label 7 by Algorithms 10 and 11.

Proof. Let $\mathcal{T}$ be the triangle spanning component that corresponds to $\triangle uvw$. By Assumption 2 and propositions 2.3.10, 2.3.12 and 2.3.15, the samples

$p \in P_{NLM}$ that witness $\mathcal{T}$ do not witness any spanning connected component $\mathcal{C} \neq \mathcal{T}$. By Assumption 2 and Algorithm 8 there is a unique connected component P_1 that witnesses $\mathcal{T}$.

As P is an ε-sample of $|X|$, and from Propositions 2.3.10 and 2.3.15, there are samples $p_u, p_v, p_w, p_{uv}, p_{vw}, p_{uw} \in P_1$ such that

$$\|p_u - u\|,\ \|p_v - v\|,\ \|p_w - w\| \leq \varepsilon,$$
$$d(\overline{uv}, p_{uv}),\ d(\overline{vw}, p_{vw}),\ d(\overline{uw}, p_{uw}) \leq \varepsilon.$$

Hence, P_1 is assigned label 7 by Algorithms 10 and 11. □

Lemma 2.4.13. *Let $\triangle uvw$ be a triangle of X, such that there is some locally maximal cell $\sigma \in X$ with $\sigma \neq \triangle uvw$, such that $v \in \sigma$, without loss of generality, and for all locally maximal $\tau \in X$, $\tau \neq \sigma, \triangle uvw$, either $\triangle uvw \cap \tau = v$ or $\triangle uvw \cap \tau = \emptyset$.*

Then, there are exactly two partitions P_1, P_2 of P_{NLM} which witness $\mathcal{T}$, where $\mathcal{T}$ is the edge spanning component corresponding to $\triangle uvw$. Further, P_1 is given label 0 and P_2 label 5 by Algorithms 10 and 11.

Proof. Let $\mathcal{T}$ be the triangle spanning component that corresponds to $\triangle uvw$. By Assumption 2 and Propositions 2.3.10 to 2.3.12 and 2.3.15, any spanning connected component $\mathcal{C}$ witnessed by samples $p \in P_{NLM}$ that witness $\mathcal{T}$ corresponds to a locally maximal cell τ such that $\triangle uvw \cap \tau \neq \emptyset$.

We need to split into two cases:

1. there is a unique locally maximal cell $\tau \in X$ with $\triangle uvw \cap \tau = v$
2. there are at least two locally maximal cells $\tau, \sigma \in X$, $\tau \neq \sigma$ with $\triangle uvw \cap \tau = \triangle uvw \cap \sigma = v$.

Case 1: We assumed there was a unique locally maximal τ with $\triangle uvw \cap \tau = v$, and hence, by Propositions 2.4.7 and 2.4.8 there is some spanning component $\mathcal{C}_\tau$ which corresponds to τ. By Assumption 2 and Propositions 2.3.10 to 2.3.12 and 2.3.15, in Algorithm 8 there is a single partition P_1 of P_{NLM} which witnesses $\mathcal{T}$ and $\mathcal{C}_\tau$, and there is a unique partition P_2 which witnesses just $\mathcal{T}$. Further, P_1 is assigned label 0 and P_2 label 5 by Algorithms 10 and 11.

Case 2: From our assumptions, there are two locally maximal cells $\tau, \sigma \in X$, $\tau \neq \sigma$ such that

$$\tau \cap \triangle uvw = v = \sigma \cap \triangle uvw.$$

By Propositions 2.4.7 and 2.4.8 there is some spanning component $\mathcal{C}_\tau$ which corresponds to τ, and some spanning component $\mathcal{C}_\sigma$ which corresponds to σ.

By Assumption 2 and from Algorithm 8, there is a single partition P_1 of P_{NLM} which witnesses $\mathcal{T}, \mathcal{C}_\tau, \mathcal{C}_\sigma$, and no partitions which witness a subset of these spanning components. This holds, by induction, for any locally maximal cell $\tau' \in X, \tau' \neq \tau, \sigma$ with $\tau' \cap \triangle uvw = v$. Similarly, there is a single partition P_2 of P_{NLM} which witnesses only $\mathcal{T}$. Further, P_1 is assigned label 0 and P_2 label 5 by Algorithms 10 and 11. □

Lemma 2.4.14. *Let $\triangle uvw$ be a triangle of X, such that there is some locally maximal cell $\sigma \in X$ with $\sigma \neq \triangle uvw$, such that $v \in \sigma$, without loss of generality, and for all locally maximal $\tau \in X$, $\tau \neq \sigma, \triangle uvw$, either $\triangle uvw \cap \tau = \overline{uv}$ or $\triangle uvw \cap \tau = \emptyset$.*

Then, there are exactly two partitions P_1, P_2 of P_{NLM} which witness $\mathcal{T}$, where $\mathcal{T}$ is the edge spanning component corresponding to $\triangle uvw$. Further, P_1 is given label 0 and P_2 label 5 by Algorithms 10 and 11.

Proof. Let $\mathcal{T}$ be the triangle spanning component that corresponds to $\triangle uvw$. By Assumption 2 and Propositions 2.3.10 to 2.3.12 and 2.3.15, any spanning connected component $\mathcal{C}$ witnessed by samples $p \in P_{NLM}$ that witness $\mathcal{T}$ corresponds to a locally maximal cell τ such that $\triangle uvw \cap \tau \neq \emptyset$.

We need to split into two cases:

1. there is a unique locally maximal cell $\tau \in X$ with $\triangle uvw \cap \tau = \overline{uv}$
2. there are at least two locally maximal cells $\tau, \sigma \in X$, $\tau \neq \sigma$ with $\triangle uvw \cap \tau = \triangle uvw \cap \sigma = \overline{uv}$.

Case 1: We assumed there was a unique locally maximal τ with $\triangle uvw \cap \tau = \overline{uv}$, and hence, by Propositions 2.4.7 and 2.4.8 there is some spanning component $\mathcal{C}_\tau$ which corresponds to τ. By Assumption 2, Propositions 2.3.10 to 2.3.12 and 2.3.15, in Algorithm 8 there is a single partition P_1 of P_{NLM} which witnesses $\mathcal{T}$ and $\mathcal{C}_\tau$, and there is a unique partition P_2

which witnesses just $\mathcal{T}$. Further, P_1 is assigned label 1 and P_2 label 4 by Algorithms 10 and 11.

Case 2: From our assumptions, there are two locally maximal cells $\tau, \sigma \in X$, $\tau \neq \sigma$ such that

$$\tau \cap \triangle uvw = \overline{uv} = \sigma \cap \triangle uvw.$$

By Propositions 2.4.7 and 2.4.8 there is some spanning component $\mathcal{C}_\tau$ which corresponds to τ, and some spanning component $\mathcal{C}_\sigma$ which corresponds to σ.

By Assumption 2 and from Algorithm 8, there is a single partition P_1 of P_{NLM} which witnesses $\mathcal{T}, \mathcal{C}_\tau, \mathcal{C}_\sigma$, and no partitions which witness a subset of these spanning components. This holds, by induction, for any locally maximal cell $\tau' \in X, \tau' \neq \tau, \sigma$ with $\tau' \cap \triangle uvw = v$. Similarly, there is a single partition P_2 of P_{NLM} which witnesses only $\mathcal{T}$. Further, P_1 is assigned label 1 and P_2 label 4 by Algorithms 10 and 11. □

Lemma 2.4.15. *Let $\triangle uvw$ be a triangle of X, such that there are some locally maximal cells $\sigma_1 \neq \sigma_2 \in X$ with $\sigma_1, \sigma_2 \neq \triangle uvw$, such that*

$$\sigma_1 \cap \triangle uvw = v$$
$$\sigma_2 \cap \triangle uvw = u$$

and for all other locally maximal cells $\tau \in X$, either

1. *$\tau \cap \triangle uvw = v$,*
2. *$\tau \cap \triangle uvw = u$,*
3. *$\tau \cap \triangle uvw = \emptyset$.*

Then, there are exactly three partitions P_1, P_2, P_2 of P_{NLM} which witness $\mathcal{T}$, where $\mathcal{T}$ is the edge spanning component corresponding to $\triangle uvw$. Further, P_1, P_2 are given label 0 and P_3 label 6 by Algorithms 10 and 11.

Proof. Let $\mathcal{T}$ be the triangle spanning component which corresponds to $\triangle uvw$. Then, the proof is an adaption of the proof of Lemma 2.4.13. By combining the arguments at the two shared vertices, there are three partitions P_1, P_2, P_3 from Algorithm 8 which witness $\mathcal{T}$, and there are spanning connected components $\mathcal{C}_1, \mathcal{C}_2$ such that P_1 witnesses $\mathcal{C}_1$ but not $\mathcal{C}_2$, and P_2 witnesses $\mathcal{C}_2$ but not $\mathcal{C}_1$. Further, P_3 only witnesses $\mathcal{T}$. Hence, P_1, P_2 are labelled with 0 and P_3 with 6. □

Lemma 2.4.16. *Let $\triangle uvw$ be a triangle of X, such that there are some locally maximal cells $\sigma_1 \neq \sigma_2 \in X$ with $\sigma_1, \sigma_2 \neq \triangle uvw$, such that*

$$\sigma_1 \cap \triangle uvw = \overline{uv}$$
$$\sigma_2 \cap \triangle uvw = v$$

and for all other locally maximal cells $\tau \in X$, either

1. *$\tau \cap \triangle uvw = \overline{uv}$,*
2. *$\tau \cap \triangle uvw = v$,*
3. *$\tau \cap \triangle uvw = \emptyset$.*

Then, there are exactly three partitions P_1, P_2, P_2 of P_{NLM} which witness $\mathcal{T}$, where $\mathcal{T}$ is the edge spanning component corresponding to $\triangle uvw$. Further, P_1 has label 0, P_2 label 1 and P_3 label 4 by Algorithms 10 and 11.

Proof. Let $\mathcal{T}$ be the triangle spanning component which corresponds to $\triangle uvw$. Then, the proof is an adaption of the proof of Lemmas 2.4.13 and 2.4.14. By combining the arguments there are three partitions P_1, P_2, P_3 from Algorithm 8 which witness $\mathcal{T}$, and there are spanning connected components $\mathcal{C}_1, \mathcal{C}_2$ such that P_1 witnesses $\mathcal{C}_1$ and $\mathcal{C}_2$, and P_2 witnesses $\mathcal{C}_2$ but not $\mathcal{C}_1$. Further, P_3 only witnesses $\mathcal{T}$. Hence, P_1 is labelled with 0, P_2 with 1 and P_3 with 3. □

Lemma 2.4.17. *Let $\triangle uvw$ be a triangle of X, such that there are some locally maximal cells $\sigma_1 \neq \sigma_2 \in X$ with $\sigma_i \neq \triangle uvw$ and $\sigma_i \neq \sigma_j$ for $i \neq j$, such that*

$$\sigma_1 \cap \triangle uvw = \overline{uv}$$
$$\sigma_2 \cap \triangle uvw = w$$

and for all other locally maximal cells $\tau \in X$, either

1. $\tau \cap \triangle uvw = \overline{uv}$,
2. $\tau \cap \triangle uvw = w$,
3. $\tau \cap \triangle uvw = \emptyset$.

Then, there are exactly three partitions P_1, P_2, P_2 of P_{NLM} which witness $\mathcal{T}$, where $\mathcal{T}$ is the edge spanning component corresponding to $\triangle uvw$. Further, P_1 has label 0, P_2 label 2 and P_3 label 9 by Algorithms 10 and 11.

Proof. Let $\mathcal{T}$ be the triangle spanning component which corresponds to $\triangle uvw$. Then, the proof is an adaption of the proof of Lemmas 2.4.13 and 2.4.14. By combining the arguments there are three partitions P_1, P_2, P_3 from Algorithm 8 which witness $\mathcal{T}$, and there are spanning connected components $\mathcal{C}_1, \mathcal{C}_2$ such that P_1 witnesses $\mathcal{C}_1$ but not $\mathcal{C}_2$, and P_2 witnesses $\mathcal{C}_2$ but not $\mathcal{C}_1$. Further, P_3 only witnesses $\mathcal{T}$. Hence, P_1 is labelled with 0, P_2 with 2 and P_3 with 9. □

Lemma 2.4.18. *Let $\triangle uvw$ be a triangle of X, such that there are some locally maximal cells $\sigma_1, \sigma_2, \sigma_3 \in X$ with $\sigma_i \neq \triangle uvw$ and $\sigma_i \neq \sigma_j$ for $i \neq j$, such that*

$$\sigma_1 \cap \triangle uvw = u$$
$$\sigma_2 \cap \triangle uvw = v$$
$$\sigma_3 \cap \triangle uvw = w$$

and for all other locally maximal cells $\tau \in X$, either

1. $\tau \cap \triangle uvw = u$,
2. $\tau \cap \triangle uvw = v$,
3. $\tau \cap \triangle uvw = w$,
4. $\tau \cap \triangle uvw = \emptyset$.

Then, there are exactly four partitions P_1, P_2, P_3, P_4 of P_{NLM} which witness $\mathcal{T}$, where $\mathcal{T}$ is the edge spanning component corresponding to $\triangle uvw$. Further, P_1, P_2 and P_3 are labelled with 0 and P_4 with 8 by Algorithms 10 and 11.

Proof. Let $\mathcal{T}$ be the triangle spanning component which corresponds to $\triangle uvw$. Then, the proof is an adaption of the proof of Lemmas 2.4.13 and 2.4.14. By combining the arguments there are three partitions P_1, P_2, P_3, P_4 from Algorithm 8 which witness $\mathcal{T}$, and there are spanning connected components $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_\ni$ such that P_1 witnesses $\mathcal{C}_1$ but not $\mathcal{C}_2, \mathcal{C}_3$, P_2 witnesses $\mathcal{C}_2$ but not $\mathcal{C}_1, \mathcal{C}_3$, and P_2 witnesses $\mathcal{C}_3$ but not $\mathcal{C}_1, \mathcal{C}_2$. Further, P_4 only witnesses $\mathcal{T}$. Hence, P_1, P_2 and P_3 are labelled with 0 and P_4 with 8. □

Lemma 2.4.19. *Let $\triangle uvw$ be a triangle of X, such that there are some locally maximal cells $\sigma_1, \sigma_2, \sigma_3 \in X$ with $\sigma_i \neq \triangle uvw$ and $\sigma_i \neq \sigma_j$ for $i \neq j$, such that*

$$\sigma_1 \cap \triangle uvw = \overline{uv}$$
$$\sigma_2 \cap \triangle uvw = v$$
$$\sigma_3 \cap \triangle uvw = w$$

and for all other locally maximal cells $\tau \in X$, either

1. *$\tau \cap \triangle uvw = \overline{uv}$,*
2. *$\tau \cap \triangle uvw = v$,*
3. *$\tau \cap \triangle uvw = w$,*
4. *$\tau \cap \triangle uvw = \emptyset$.*

Then, there are exactly four partitions P_1, P_2, P_3, P_4 of P_{NLM} which witness $\mathcal{T}$, where $\mathcal{T}$ is the edge spanning component corresponding to $\triangle uvw$. Further, P_1 is labelled with 1, P_2, P_3 with 0 and P_4 with 9 by Algorithms 10 and 11.

Proof. Let $\mathcal{T}$ be the triangle spanning component which corresponds to $\triangle uvw$. Then, the proof is an adaption of the proof of Lemmas 2.4.13 and 2.4.14. By combining the arguments there are three partitions

$$P_1, P_2, P_3, P_4$$

from Algorithm 8 which witness $\mathcal{T}$, and there are spanning connected components $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_\ni$ such that P_1 witnesses $\mathcal{C}_1$ but not $\mathcal{C}_2, \mathcal{C}_3$, P_2 witnesses

$\mathcal{C}_1, \mathcal{C}_2$ but not $\mathcal{C}_3$, and P_2 witnesses $\mathcal{C}_3$ but not $\mathcal{C}_1, \mathcal{C}_2$. Further, P_4 only witnesses $\mathcal{T}$. Hence, P_1, P_2 and P_3 are labelled with 0 and P_4 with 8. □

Lemma 2.4.20. *Let $\triangle uvw$ be a triangle of X, such that there are some locally maximal cells $\sigma_1, \sigma_2, \sigma_3 \in X$ with $\sigma_i \neq \triangle uvw$ and $\sigma_i \neq \sigma_j$ for $i \neq j$, such that*

$$\sigma_1 \cap \triangle uvw = \overline{uv}$$
$$\sigma_2 \cap \triangle uvw = u$$
$$\sigma_3 \cap \triangle uvw = v$$

and for all other locally maximal cells $\tau \in X$, either

1. *$\tau \cap \triangle uvw = \overline{uv}$,*
2. *$\tau \cap \triangle uvw = u$,*
3. *$\tau \cap \triangle uvw = v$,*
4. *$\tau \cap \triangle uvw = \emptyset$.*

Then, there are exactly four partitions P_1, P_2, P_3, P_4 of P_{NLM} which witness $\mathcal{T}$, where $\mathcal{T}$ is the edge spanning component corresponding to $\triangle uvw$. Further, P_1 is labelled with 3, P_2, P_3 with 0 and P_4 with 4 by Algorithms 10 and 11.

Proof. Let $\mathcal{T}$ be the triangle spanning component which corresponds to $\triangle uvw$. Then, the proof is an adaption of the proof of Lemmas 2.4.13 and 2.4.14. By combining the arguments there are three partitions P_1, P_2, P_3, P_4 from Algorithm 8 which witness $\mathcal{T}$, and there are spanning connected components $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_\ni$ such that P_1 witnesses $\mathcal{C}_1$ but not $\mathcal{C}_2, \mathcal{C}_3$, P_2 witnesses $\mathcal{C}_1, \mathcal{C}_2$ but not $\mathcal{C}_3$, and P_2 witnesses $\mathcal{C}_1, \mathcal{C}_3$ but not $\mathcal{C}_2$. Further, P_4 only witnesses $\mathcal{T}$. Hence, P_1 is labelled with 3, P_2, P_3 with 0 and P_4 with 4. □

Lemma 2.4.21. *Let $\triangle uvw$ be a triangle of X, such that there are some locally maximal cells $\sigma_1, \sigma_2, \sigma_3 \in X$ with $\sigma_i \neq \triangle uvw$ and $\sigma_i \neq \sigma_j$ for $i \neq j$, such that*

$$\sigma_1 \cap \triangle uvw = \overline{uv}$$
$$\sigma_2 \cap \triangle uvw = \overline{vw}$$
$$\sigma_3 \cap \triangle uvw = v$$

and for all other locally maximal cells $\tau \in X$, either

1. $\tau \cap \triangle uvw = \overline{uv}$*,*
2. $\tau \cap \triangle uvw = \overline{vw}$*,*
3. $\tau \cap \triangle uvw = v$*,*
4. $\tau \cap \triangle uvw = \emptyset$*.*

Then, there are exactly four partitions P_1, P_2, P_3, P_4 of P_{NLM} which witness $\mathcal{T}$, where $\mathcal{T}$ is the edge spanning component corresponding to $\triangle uvw$. Further, P_1 is labelled with 0, P_2, P_3 with 1, and P_3 with 3 by Algorithms 10 and 11.

Proof. Let $\mathcal{T}$ be the triangle spanning component which corresponds to $\triangle uvw$. Then, the proof is an adaption of the proof of Lemmas 2.4.13 and 2.4.14. By combining the arguments there are four partitions P_1, P_2, P_3, P_4 from Algorithm 8 which witness $\mathcal{T}$, and there are spanning connected components $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_\ni$ such that P_1 witnesses $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$, P_2 witnesses $\mathcal{C}_1, \mathcal{C}_2$ but not $\mathcal{C}_3$, and P_3 witnesses $\mathcal{C}_1, \mathcal{C}_3$ but not $\mathcal{C}_2$. Further, P_4 only witnesses $\mathcal{T}$. Hence, P_1 is labelled with 0, P_2, P_3 with 1, and P_3 with 3. □

Lemma 2.4.22. *Let $\triangle uvw$ be a triangle of X, such that there are some locally maximal cells $\sigma_1, \sigma_2, \sigma_3, \sigma_4 \in X$ with $\sigma_i \neq \triangle uvw$ and $\sigma_i \neq \sigma_j$ for $i \neq j$, such that*

$$\sigma_1 \cap \triangle uvw = u$$
$$\sigma_2 \cap \triangle uvw = v$$
$$\sigma_3 \cap \triangle uvw = w$$
$$\sigma_4 \cap \triangle uvw = \overline{uv}$$

and for all other locally maximal cells $\tau \in X$, either

1. $\tau \cap \triangle uvw = u$,
2. $\tau \cap \triangle uvw = v$,
3. $\tau \cap \triangle uvw = w$,
4. $\tau \cap \triangle uvw = \overline{uv}$
5. $\tau \cap \triangle uvw = \emptyset$.

Then, there are exactly five partitions P_1, P_2, P_3, P_4, P_5 of P_{NLM} which witness $\mathcal{T}$, where $\mathcal{T}$ is the edge spanning component corresponding to $\triangle uvw$. Further, P_1, P_2, P_3 are labelled with 0, and P_4 with 8 by Algorithms 10 and 11.

Proof. Let $\mathcal{T}$ be the triangle spanning component which corresponds to $\triangle uvw$. Then, the proof is an adaption of the proof of Lemmas 2.4.13 and 2.4.14. By combining the arguments there are four partitions P_1, P_2, P_3, P_4 from Algorithm 8 which witness $\mathcal{T}$, and there are spanning connected components $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_\ni, \mathcal{C}_\ni$ such that P_1 witnesses $\mathcal{C}_1$ and not $\mathcal{C}_2, \mathcal{C}_3$, P_2 witnesses $\mathcal{C}_1$ and not $\mathcal{C}_2, \mathcal{C}_3$, and P_3 witnesses $\mathcal{C}_3$ but not $\mathcal{C}_1, \mathcal{C}_2$. Further, P_4 only witnesses $\mathcal{T}$. Hence, P_1, P_2, P_3 are labelled with 0, and P_4 with 8. □

Lemma 2.4.23. *Let $\triangle uvw$ be a triangle of X, such that there are some locally maximal cells $\sigma_1, \sigma_2, \sigma_3, \sigma_4 \in X$ with $\sigma_i \neq \triangle uvw$ and $\sigma_i \neq \sigma_j$ for $i \neq j$, such that*

$$\sigma_1 \cap \triangle uvw = u$$
$$\sigma_2 \cap \triangle uvw = v$$
$$\sigma_3 \cap \triangle uvw = \overline{vw}$$
$$\sigma_4 \cap \triangle uvw = \overline{uv}$$

and for all other locally maximal cells $\tau \in X$, either

1. $\tau \cap \triangle uvw = u$,
2. $\tau \cap \triangle uvw = v$,
3. $\tau \cap \triangle uvw = w$,
4. $\tau \cap \triangle uvw = \overline{uv}$
5. $\tau \cap \triangle uvw = \overline{vw}$
6. $\tau \cap \triangle uvw = \emptyset$.

Then, there are exactly five partitions P_1, P_2, P_3, P_4, P_5 of P_{NLM} which witness $\mathcal{T}$, where $\mathcal{T}$ is the edge spanning component corresponding to $\triangle uvw$. Further, P_1, P_2 are labelled with 0, P_3 with 1, and P_4, P_5 with 3 by Algorithms 10 and 11.

Proof. Let $\mathcal{T}$ be the triangle spanning component which corresponds to $\triangle uvw$. Then, the proof is an adaption of the proof of Lemmas 2.4.13 and 2.4.14. By combining the arguments, there are five partitions

$$P_1, P_2, P_3, P_4, P_5$$

from Algorithm 8 which witness $\mathcal{T}$, and there are spanning connected components $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_\ni, \mathcal{C}_\ni, \mathcal{C}_\triangle$ such that P_1 witnesses $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_\triangle$ and not $\mathcal{C}_3$, P_2 witnesses $\mathcal{C}_2$ and not $\mathcal{C}_1, \mathcal{C}_3, \mathcal{C}_4$, P_3 witnesses $\mathcal{C}_2, \mathcal{C}_3$ but not $\mathcal{C}_1, \mathcal{C}_4$, and P_4 witnesses $\mathcal{C}_4$ but not $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$. Further, P_5 only witnesses $\mathcal{T}$, and hence P_4 only witnesses $\mathcal{T}$. Hence, P_1, P_2 are labelled with 0, P_3 with 1, and P_4, P_5 with 3. □

Lemma 2.4.24. *Let $\triangle uvw$ be a triangle of X, such that there are some locally maximal cells $\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5 \in X$ with $\sigma_i \neq \triangle uvw$ and $\sigma_i \neq \sigma_j$ for $i \neq j$, such that*

$$\begin{aligned}
\sigma_1 \cap \triangle uvw &= u \\
\sigma_2 \cap \triangle uvw &= v \\
\sigma_3 \cap \triangle uvw &= w \\
\sigma_4 \cap \triangle uvw &= \overline{uv} \\
\sigma_5 \cap \triangle uvw &= \overline{vw}
\end{aligned}$$

and for all other locally maximal cells $\tau \in X$, either

1. $\tau \cap \triangle uvw = u$,
2. $\tau \cap \triangle uvw = v$,
3. $\tau \cap \triangle uvw = w$,
4. $\tau \cap \triangle uvw = \overline{uv}$
5. $\tau \cap \triangle uvw = \overline{vw}$

6. $\tau \cap \triangle uvw = \emptyset$.

Then, there are exactly five partitions $P_1, P_2, P_3, P_4, P_5, P_6$ *of* P_{NLM} *which witness* $\mathcal{T}$*, where* $\mathcal{T}$ *is the edge spanning component corresponding to* $\triangle uvw$*. Further,* P_1, P_2, P_3 *are labelled with* 0*,* P_4, P_5, P_6 *with* 3 *by Algorithms 10 and 11.*

Proof. Let $\mathcal{T}$ be the triangle spanning component which corresponds to $\triangle uvw$. Then, the proof is an adaption of the proof of Lemmas 2.4.13 and 2.4.14. By combining the arguments there are six partitions

$$P_1, P_2, P_3, P_4, P_5, P_6$$

from Algorithm 8 which witness $\mathcal{T}$, and there are spanning connected components $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4, \mathcal{C}_5$ such that P_1 witnesses $\mathcal{C}_1, \mathcal{C}_4$ and not $\mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_5$, P_2 witnesses $\mathcal{C}_2, \mathcal{C}_4, \mathcal{C}_5$ and not $\mathcal{C}_1, \mathcal{C}_3$, P_3 witnesses $\mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_5$ but not $\mathcal{C}_1, \mathcal{C}_4$, P_4 witnesses $\mathcal{C}_4$ but not $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_5$, and P_5 witnesses $\mathcal{C}_5$ but not $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4$. Further, P_6 only witnesses $\mathcal{T}$, and hence P_1, P_2, P_3 are labelled with 0, P_4, P_5, P_6 with 3. □

Lemma 2.4.25. *Let* $\triangle uvw$ *be a triangle of* X*, such that there are some locally maximal cells* $\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5, \sigma_6 \in X$ *with* $\sigma_i \neq \triangle uvw$ *and* $\sigma_i \neq \sigma_j$ *for* $i \neq j$*, such that*

$$\sigma_1 \cap \triangle uvw = u$$

$$\sigma_2 \cap \triangle uvw = v$$

$$\sigma_3 \cap \triangle uvw = w$$

$$\sigma_4 \cap \triangle uvw = \overline{uv}$$

$$\sigma_5 \cap \triangle uvw = \overline{vw}$$

$$\sigma_6 \cap \triangle uvw = \overline{uw}$$

and for all other locally maximal cells $\tau \in X$*, either*

1. $\tau \cap \triangle uvw = u$,
2. $\tau \cap \triangle uvw = v$,
3. $\tau \cap \triangle uvw = w$,
4. $\tau \cap \triangle uvw = \overline{uv}$,

5. $\tau \cap \triangle uvw = \overline{vw}$,
6. $\tau \cap \triangle uvw = \overline{uw}$,
7. $\tau \cap \triangle uvw = \emptyset$.

Then, there are exactly six partitions $P_1, P_2, P_3, P_4, P_5, P_6$ of P_{NLM} which witness $\mathcal{T}$, where $\mathcal{T}$ is the edge spanning component corresponding to $\triangle uvw$. Further, P_1, P_2, P_3 are labelled with 0, P_4, P_5, P_6 with 3 by by Algorithms 10 and 11.

Proof. Let $\mathcal{T}$ be the triangle spanning component which corresponds to $\triangle uvw$. Then, the proof is an adaption of the proof of Lemmas 2.4.13 and 2.4.14. By combining the arguments there are six partitions

$$P_1, P_2, P_3, P_4, P_5, P_6$$

from Algorithm 8 which witness $\mathcal{T}$, and there are spanning connected components $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4, \mathcal{C}_5, \mathcal{C}_6$ such that P_1 witnesses $\mathcal{C}_1, \mathcal{C}_4, \mathcal{C}_6$ and not $\mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_5$, P_2 witnesses $\mathcal{C}_2, \mathcal{C}_4, \mathcal{C}_5$ and not $\mathcal{C}_1, \mathcal{C}_3, \mathcal{C}_6$, P_3 witnesses $\mathcal{C}_3, \mathcal{C}_5, \mathcal{C}_6$ but not $\mathcal{C}_1, \mathcal{C}_2\mathcal{C}_4$, P_4 witnesses $\mathcal{C}_4$ but not $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_5, \mathcal{C}_6$, and P_5 witnesses $\mathcal{C}_5$ but not $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4, \mathcal{C}_6$, and P_6 witnesses $\mathcal{C}_6$ but not $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4, \mathcal{C}_5$. Hence P_1, P_2, P_3 are labelled with 0, P_4, P_5, P_6 with 3. □

Theorem 2.4.26. *Let P be an ε-sample of an embedded 2-complex $|X| \subset \mathbb{R}^n$ satisfying Assumption 2, and let B be the graph obtained from Algorithm 12.*

Then, we can complete B to be the incidence graph of X, to recover the abstract structure.

Proof. From Propositions 2.4.6 to 2.4.8, we correctly identify the locally maximal components of X. It remains to show that we correctly learn the number of not locally maximal cells, and the incidence relationship.

For a locally maximal edge, we need to identify two vertices as its faces. To do so, we must identify which partition(s) of P_{NLM} correspond to these vertices.

Take a spanning edge component $\mathcal{E}$. Then there is some locally maximal edge $\overline{uv}$ corresponding to $\mathcal{E}$. There are two cases to consider:

A: $\overline{uv}$ is disconnected from every other part of X,
B: $\overline{uv}$ is not disconnected every other part of X.

Case A: From Propositions 2.3.10 to 2.3.12, 2.3.14 and 2.3.15 and Assumption 2, there is a single partition $P_i \subset P_{NLM}$ which contains points p such that $\mathcal{E} \cap B_{(R+\varepsilon)/\kappa+3\varepsilon}(p) \neq \emptyset$. Hence, P_i contains samples p such that either $\|v-p\| \leq \frac{3R}{2}+\varepsilon$ or $\|u-p\| \leq \frac{3R}{2}+\varepsilon$, and P_i corresponds to u and v. In this case, P_i is labelled with -1 in Algorithm 10. This occurs only when $\overline{uv}$ is disconnected from the rest of $|X|$; hence, we infer the two boundary vertices.

Case B: As $\overline{uv}$ is not disconnected, there is some locally maximal cell $\sigma \in X$, $\sigma \neq \overline{uv}$ such that either u or v is a vertex of σ. Without loss of generality, let $v \in \sigma$. For the vertices u and v let the set of locally maximal faces they see be $S(u)$ and $S(v)$, respectively. As X is a 2-complex, and $\overline{uv}$ a locally maximal edge, $\sigma \notin S(u)$. Hence, there are two partitions, P_u, P_v, which correspond to the vertices u and v, respectively. In this case, P_u and P_v are labelled with 0 in Algorithm 10.

We now need to examine how we identify the faces of triangles.

For a triangle spanning component $\mathcal{T}$, let $\mathcal{P}_{\mathcal{T}}$ be the set of partitions P_i of P_{NLM} such that $d(\mathcal{T}, P_i) \leq 3\varepsilon$. There are a few cases we need to consider to ensure we correctly recover the structure of X:

1. $|\mathcal{P}_{\mathcal{T}}|= 1$,
2. $|\mathcal{P}_{\mathcal{T}}|= 2$,
3. $|\mathcal{P}_{\mathcal{T}}|= 3$,
4. $|\mathcal{P}_{\mathcal{T}}|= 4$,
5. $|\mathcal{P}_{\mathcal{T}}|= 5$,
6. $|\mathcal{P}_{\mathcal{T}}|= 6$.

Let the weight 2 node labelled with $\mathcal{E}$ be t.

Case 1 $|\mathcal{P}_{\mathcal{T}}|= 1$: Let P_1 be the single partition in $\mathcal{P}_{\mathcal{T}}$.

This can only occur if the triangle $\triangle uvw$ corresponding to $\mathcal{T}$ does not share any faces with another cell. Then, P_1 corresponds to three edges and three vertices and is correctly labelled with 7 by Algorithms 10 and 11. Let the corresponding weight 1 nodes of B be e_1, e_2, e_3 and the weight 0 nodes be v_1, v_2, v_3. We add an edge between t and $e_1, e_2, e_3, v_1, v_2, v_3$ and between the following pairs:

$$(e_1, v_1), (e_1, v_2), (e_2, v_2), (e_2, v_3), (e_3, v_3), (e_3, v_1).$$

Case 2 $|\mathcal{P}_{\mathcal{T}}|=2$: Let $\mathcal{P}_{\mathcal{T}} = \{P_1, P_2\}$.

This can only occur if the triangle $\triangle uvw$ corresponding to $\mathcal{T}$ shares either a vertex, or an edge and two vertices with other triangles or locally maximal edges. Thus, either P_1 is labelled with 0 and P_2 with 5, or P_1 is labelled with 2 and P_2 with 4 by Algorithms 10 and 11.

If P_1 has label 0 and P_2 has label 5, we find the weight 0 node v_1 with label P_1 and the three weight 1 nodes e_1, e_2, e_3 and two weight 0 nodes v_2, v_3 with label P_2. Then, we add an edge between t and each of $e_1, e_2, e_3, v_1, v_2, v_3$ and between the following pairs:

$$(e_1, v_1), (e_1, v_2), (e_2, v_2), (e_2, v_3), (e_3, v_3), (e_3, v_1).$$

If P_1 has label 2 and P_2 has label 4, we find the weight 1 note e_1 and two weight 0 node v_1, v_2 with label P_1, the two weight 1 nodes e_2, e_3 and one weight 0 nodes v_3 with label P_2. We add an edge between t and each of $e_1, e_2, e_3, v_1, v_2, v_3$ and between the following pairs:

$$(e_1, v_1), (e_1, v_2), (e_2, v_2), (e_2, v_3), (e_3, v_3), (e_3, v_1).$$

Case 3 $|\mathcal{P}_{\mathcal{T}}|=3$: Let $\mathcal{P}_{\mathcal{T}} = \{P_1, P_2, P_3\}$.

This can only occur if the triangle $\triangle uvw$ corresponding to $\mathcal{T}$ shares either two vertices, or two vertices and an edge with other triangles or locally maximal edges. Thus, either P_1 and P_2 are labelled with 0 and P_2 with 6; or P_1 is labelled with 0, P_2 with 1 and P_3 with 4; or P_1 is labelled 0, P_2 with 2 and P_3 with 9 by Algorithms 10 and 11.

If P_1, P_2 have label 0 and P_3 has label 6, we find the weight 0 node v_1 with label P_1, the weight 0 node v_2 with label P_2, the three weight 1 nodes e_1, e_2, e_3 and the weight 0 node v_3 with label P_3. Then add an edge between t and each of $e_1, e_2, e_3, v_1, v_2, v_3$ and between following pairs:

$$(e_1, v_1), (e_1, v_2), (e_2, v_2), (e_2, v_3), (e_3, v_3), (e_3, v_1).$$

If P_1 has label 0, P_2 label 1 and P_3 label 4, we find the weight 0 node v_1 with label P_1, the weight 0 node v_2 with label P_2, weight 1 node e_1 with label P_2, the weight 0 node v_3 with label P_3, and the two weight 1 nodes e_2, e_3

with label P_3. Then add an edge between t and each of $e_1, e_2, e_3, v_1, v_2, v_3$ and between the following pairs:

$$(e_1, v_1), (e_1, v_2), (e_2, v_2), (e_2, v_3), (e_3, v_3), (e_3, v_1).$$

If P_1 has label 0, P_2 label 2 and P_3 label 9, we find the weight 0 node v_1 with label P_1, the weight 0 node v_2 and weight 1 node e_1 with label P_2, and the weight 1 nodes e_2, e_3 and weight 0 node v_3 with label P_3. Then add an edge between t and each of $e_1, e_2, e_3, v_1, v_2, v_3$ and between the following pairs:

$$(e_1, v_1), (e_1, v_2), (e_2, v_2), (e_2, v_3), (e_3, v_3), (e_3, v_1).$$

Case 4 $|\mathcal{P}_\mathcal{T}| = 4$: Let $\mathcal{P}_\mathcal{T} = \{P_1, P_2, P_3, P_4\}$.

This can only occur if the triangle $\triangle uvw$ corresponding to $\mathcal{T}$ shares three vertices, or three vertices and an edge, or three vertices and two edges with other triangles or locally maximal edges. Thus, either P_1, P_2 and P_3 are labelled with 0 and P_4 with 8; or P_1 is labelled with 1, P_2, P_3 with 0 and P_3 with 9; or P_1 with 3, P_2, P_3 with 0 and P_4 with 4; or P_1 is labelled with 0, P_2, P_3 with 1, and P_3 with 3 by Algorithms 10 and 11.

If P_1, P_2, P_3 have label 0 and P_4 has label 8, find the weight 0 node v_1 with label P_1, weight 0 node v_2 with label P_2, weight 0 node v_3 with label P_3, and the three weight 1 nodes e_1, e_2, e_3 with label P_4. Then add an edge between t and each of $e_1, e_2, e_3, v_1, v_2, v_3$ and between the following pairs:

$$(e_1, v_1), (e_1, v_2), (e_2, v_2), (e_2, v_3), (e_3, v_3), (e_3, v_1).$$

If P_1 has label 1, P_2, P_3 have label 0, and P_4 has label 9, find the weight 0 node v_1 and weight 1 node e_1 with label P_1, weight 0 node v_2 with label P_2, weight 0 node v_3 with label P_3, and the two weight 1 nodes e_2, e_3 with label P_4. Then add an edge between t and each of $e_1, e_2, e_3, v_1, v_2, v_3$ and between the following pairs:

$$(e_1, v_1), (e_1, v_2), (e_2, v_2), (e_2, v_3), (e_3, v_3), (e_3, v_1).$$

If P_1 has 3, P_2, P_3 label 0 and P_4 label 4;, find the weight 1 node e_1 with label P_1, weight 0 node v_1 with label P_2, weight 0 node v_2 with label P_3, and the two weight 1 nodes e_2 and weight 0 node e_3 with label P_4. Then add an edge between t and each of $e_1, e_2, e_3, v_1, v_2, v_3$ and between the following pairs:

$$(e_1, v_1), (e_1, v_2), (e_2, v_2), (e_2, v_3), (e_3, v_3), (e_3, v_1).$$

If P_1 has label 0, P_2, P_3 have label 1, and P_4 has label 3, find the weight 0 node v_1 with label P_1, weight 0 node v_2 and weight 1 node e_1 with label P_2, weight 0 node v_3 and weight 1 node e_3 with label P_3, and the two weight 1 nodes e_2 with label P_4. Then add an edge between t and each of $e_1, e_2, e_3, v_1, v_2, v_3$ and between the following pairs:

$$(e_1, v_1), (e_1, v_2), (e_2, v_2), (e_2, v_3), (e_3, v_3), (e_3, v_1).$$

Case 5 $|\mathcal{P}_\mathcal{T}|= 5$: Let $\mathcal{P}_\mathcal{T} = \{P_1, P_2, P_3, P_4, P_5\}$.

This can occur if the triangle $\triangle uvw$ corresponding to $\mathcal{T}$ shares three vertices and two edges; or three vertcies and one edge with other triangles or locally maximal edges. Thus, P_1, P_2 are labelled with 0, P_3 with 1 and P_4, P_5 with 3; or P_1, P_2, P_3 are labelled with 0, P_4 with 3 and P_5 with 9 by Algorithms 10 and 11.

If P_1, P_2 are labelled with 0, P_3 with 1 and P_4, P_5 with 3 we find the weight 0 node v_1 with label P_1, find the weight 0 node v_2 with label P_2, find the weight 1 node e_1 and weight 0 node v_3 with label P_3, find the two weight 1 nodes e_2, e_3 with label P_4, and the two weight 1 nodes e_2, e_3 with label P_5. Then add an edge between t and each of $e_1, e_2, e_3, v_1, v_2, v_3$ and between e_i with label P_i and v_j with label P_j if $d(P_i, P_j)$.

If P_1, P_2, P_3 are labelled with 0, P_4 with 3 and P_5 with 9 we find the weight 0 node v_1 with label P_1, find the weight 0 node v_2 with label P_2, find the weight 0 node v_3 with label P_3, find the weight 1 node e_1 with label P_4, and the two weight 1 nodes e_2, e_3 with label P_5. Then add an edge between t and each of $e_1, e_2, e_3, v_1, v_2, v_3$ and between e_i with label P_i and v_j with label P_j if $d(P_i, P_j)$.

Case 6 $|\mathcal{P}_\mathcal{T}|= 6$: Let $\mathcal{P}_\mathcal{T} = \{P_1, P_2, P_3, P_4, P_5, P_6\}$.

This can only occur if the triangle $\triangle uvw$ corresponding to $\mathcal{T}$ shares three vertices and two edges, or three vertices and three edges with other triangles or locally maximal edges. In either case, P_1, P_2, P_3 are labelled with 0, P_4, P_5, P_6 with 3 by Algorithms 10 and 11.

So we find the weight 0 node v_1 with label P_1, find the weight 0 node v_2 with label P_2, find the weight 0 node v_3 with label P_3, find the weight 1 node e_1 with label P_4, the weight 1 node e_2 with label P_5, and the weight 1 node e_3 with label P_6. Then add an edge between t and each of $e_1, e_2, e_3, v_1, v_2, v_3$ and between e_i with label P_i and v_j with label P_j if $d(P_i, P_j)$.

In each of these 6 cases, we have connected the weight 2 node t corresponding to the cell τ to each weight 1 node e corresponding to an edge σ_e of τ, as well as to each weight 0 node v corresponding to a vertex σ_v of τ. Further, in the process, we also connect the weight 1 node e and weight 0 node v if σ_v is a vertex of σ_e.

We have shown that the weight 2 nodes of B correspond bijectively to the triangles of X, the weight 1 nodes of B correspond bijectively to the edges of X, and the weight 0 nodes of B correspond bijectively to the vertices of X. We have also shown that for any pair of nodes n_1, n_2 with corresponding cells σ_1, σ_2, there an edge between them if and only if $\sigma_1 \subset \sigma_2$ or $\sigma_2 \subset \sigma_1$.

Hence, B is the incidence graph of X. □

In this chapter, we have presented a method for learning the abstract structure X underlying an embedded 2-simplicial complex $|X|= (X, \Theta)$ (satisfying Assumption 2) from an ε-sample P. In Chapter 1, we also presented an algorithm for modelling the linear embedding of a graph. For abstract 2-complexes, modelling the embedding is future work. In particular, to model embeddings that are not linear or where we allow for cells of dimension 2, which are not triangles (along the lines of CW-complexes), we need to develop the process for learning the faces of locally maximal cells further.

CHAPTER 3

Classification of Human Mesenchymal Stem Cells

> Now is a time for simplicity.
> ...for, dare I say it, kindness.
>
> Margaret Edson, *W;t* (Vivian Bearing Scene 12)

Cell biology relies heavily on microscopy methods to visualise the specimens on various length scales from the tissue level to single cells and down to the sub-cellular structures, eventually reaching single molecules. Despite the tremendous technological advances in the last decades (for example, the advent of super-resolution methods), methods for effectively analysing data sets with an increasing number of cells are underdeveloped.

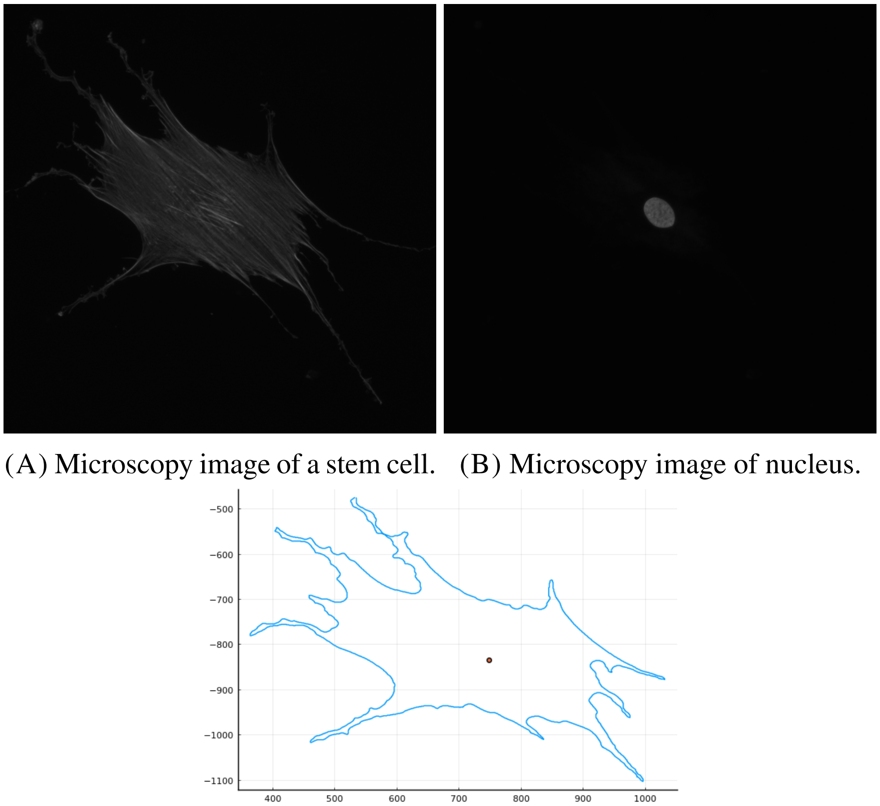
In this chapter, we present a method for quantitative analysis of the shape of cultured stem cells and use it to analyse populations of these cultured cells. The end goal of our analysis is to identify varying growth patterns in experiments, due to the mechano-response of the cells to their microenvironment. Identifying subpopulations of cells is important, as quantitative analysis of experiments with biological cells faces several problems, including cell populations may be non-homogenous, and subsets of cells may behave 'abnormally' due to either external or internal reasons (mutations, stress, substrate impurities) [**7, 11, 19**]. We use persistent homology to obtain a summary of the morphological features of biological cells. For a pair of cells, we can use this summary to compare their growth patterns and obtain a 'measure' of their (dis)similarity. Using this (dis)similarity for each pair of cells, we can identify subpopulations that exhibit similar growth patterns. There are many different ways of obtaining a summary, and computing a similarity score. In this chapter, we use the persistence of sub-level sets (Definition 3.1.1) of a radial distance function and then compute the

2-Wasserstein distance (Definition 3.1.7) between the *persistence diagrams* (Definition 3.1.3).

The mathematical methods presented in this chapter can be used to analyse any data which can be represented by closed (simple) curves in $\mathbb{R}^2$. For example, in [**17, 23, 24**], variations of persistent homology are used to identify sub-populations of animal bones, biological leaves and fonts.

Stem cells are objects which live in $\mathbb{R}^3$, but when cultured in experimental conditions, their 'thickness' is negligible. Hence most analysis is performed with 2-dimensional projections. Thus, we think of these cells as objects in $\mathbb{R}^2$, and in particular, we think of them as closed simply connected domains. An important geometric feature of these objects is their boundary, or *contour*. Currently, our analysis focuses on the contour of each cell. The task of extracting the contour of a cell from a microscopy image is challenging and scribed in [**15**]. We use the accompanying software FilamentSensor to perform the image processing and extract the contour of each cell. We extract the 'centre' of the cell from the aligned microscopy image of the nucleus. This is then used as the anchor point for a radial filtration of the contour. See Figure 3.1.

In this study, we examine the growth patterns of human mesenchymal stem cells cultured on glass. Human mesenchymal stem cells (hMSCs) are primary cells (cells isolated from live tissue) that can be collected in a variety of methods. When collected from bone marrow, it is well known and accepted that the sample will contain a small sub-population (roughly 5%) which are other primary cells from the bone marrow [**11, 19**]. Simultaneously, there is an ongoing debate about the biological nature of the population of human mesenchymal stem cells and their similarity to bone marrow fibroblasts, making it hard to distinguish the main population of hMSCs from the sub-population, which contains bone marrow fibroblasts [**11, 19**]. We examine the morphological features of the cultered cells containing human mesenchymal stem cells, bone marrow fibroblasts and other primary cells, from microscopy images of fixed and immuno-stained cells to identify and classify these populations.



(A) Microscopy image of a stem cell. (B) Microscopy image of nucleus.

(C) Plot of contour and the centre of the mass of the cell indicated in orange.

FIGURE 3.1. Example pf the microscopy image of the cell and nucleus, with a plot of the contour and the centre of the cell.

3.1. Persistent Homology on Graphs

Given a microscopy image of a fixed and immuno-stained cell, we use a graph G to represent the boundary in 2 dimensions. This graph is a list of ordered vertices, V, with edges, E, between neighbouring vertices. Note that G is connected and every vertex has degree 2, so G consists of precisely one cycle. We extract morphological information using the persistence of connected components of the sub-level sets of a radial function from the centroid of the nucleus.

For a graph G, we say two vertices v_1, v_2 are in the *same connected component* if there is a path γ from v_1 to v_2. For each connected component of G, we choose a representative vertex v and denote the set of vertices v' connected to v by $[v]$. We call the set $\{\,[v]$ for $v \in G\}$ the *connected components* of G.

To use persistent homology on G, we need to define a filtration on G. We begin by defining a family of filtered graphs.

Definition 3.1.1 (Sub-level Set)**.** *Let f be a function from a graph G to $\mathbb{R}$, and fix $a \in \mathbb{R}$. The* sub-level set $G_a := f^{-1}((-\infty, a])$ *is the subset V_a of vertices v with $f(v) \leq a$ and the set of edges E_a between any pair of neighbouring vertices which are both in V_a. Note that for any*

$$a \leq b \in \mathbb{R}$$

we have

$$f^{-1}((-\infty, a]) \subseteq f^{-1}((-\infty, b]),$$

and the sub-level sets form a sequence of nested graphs.

Remark 3.1.2. The above definition of sub-level sets is cell-wise constant, rather than piecewise-linear one. The distance of a point on an edge to the centre of the function is not the standard Euclidean distance in $\mathbb{R}^2$, but instead the maximum of the distances of the two vertices. This is not an issue, as the difference in these two values is bounded.

Given a nested sequence of graphs $G_0 \subseteq G_1 \subseteq \ldots \subseteq G_\alpha$, we can examine the changes in connected components of the graphs as we progress along the sequence. Consider some G_β, and let $C_\beta := \left\{[v_j]^\beta\right\}_{j=1}^{n_i}$ be the set of connected components in G_β. We say a connected component $[v_j]$ is *born* at time β if no vertex in $[v_j]$ it is in $C_{\beta-1}$. We say $[v_j]$ *dies* at γ where $[v_j]$ becomes path connected to a vertex born before γ. For any pair $\beta \leq \gamma$ we can define a map $\mathfrak{A}_\beta^\gamma : C_\beta \to C_\gamma$. In the current setting, once a connected component appears in G_β, it is either present in G_γ for all $\beta \leq \gamma$ or it merges with some other connected in G_γ for some $\gamma > \beta$.

Definition 3.1.3 (Persistence Diagram)**.** *Let f be a function from a graph G to $\mathbb{R}$, and let $\mathfrak{G} = \{G_a\}_{a\in\mathbb{R}}$. Let $C = \bigcup_{a\in\mathbb{R}} C_a$ be the set of connected components across the sequence of graphs $\mathfrak{G}$. The* persistence diagram*, $\mathfrak{D}(\mathfrak{G})$ of $\mathfrak{G}$ is the multi-set of points $(b_j, d_j) \in \mathbb{R}^2$, where b_j is the birth time of $[v_j] \in C$, and d_j its death time.*

We can also define these filtrations, and persistence diagrams algebraically, including persistence modules, as in [9].

Example 3.1.4. Let G be the following graph embedded in $\mathbb{R}^2$.

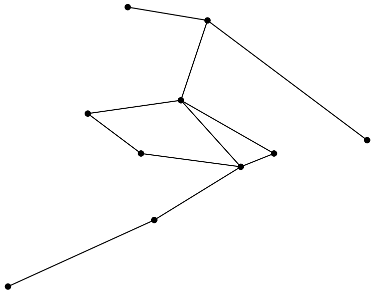


FIGURE 3.2. Graph G embedded in $\mathbb{R}^2$ for Example 3.1.4.

Consider the filtration in Figure 3.3, where we have set $G_0 = \emptyset$.

Using integer steps as our filtration parameter, we obtain the persistence diagrams in Figure 3.4.

The persistence diagram provides a summary of the changes in the connected components as we progress along the sequence of graphs. Given two sequences of graphs

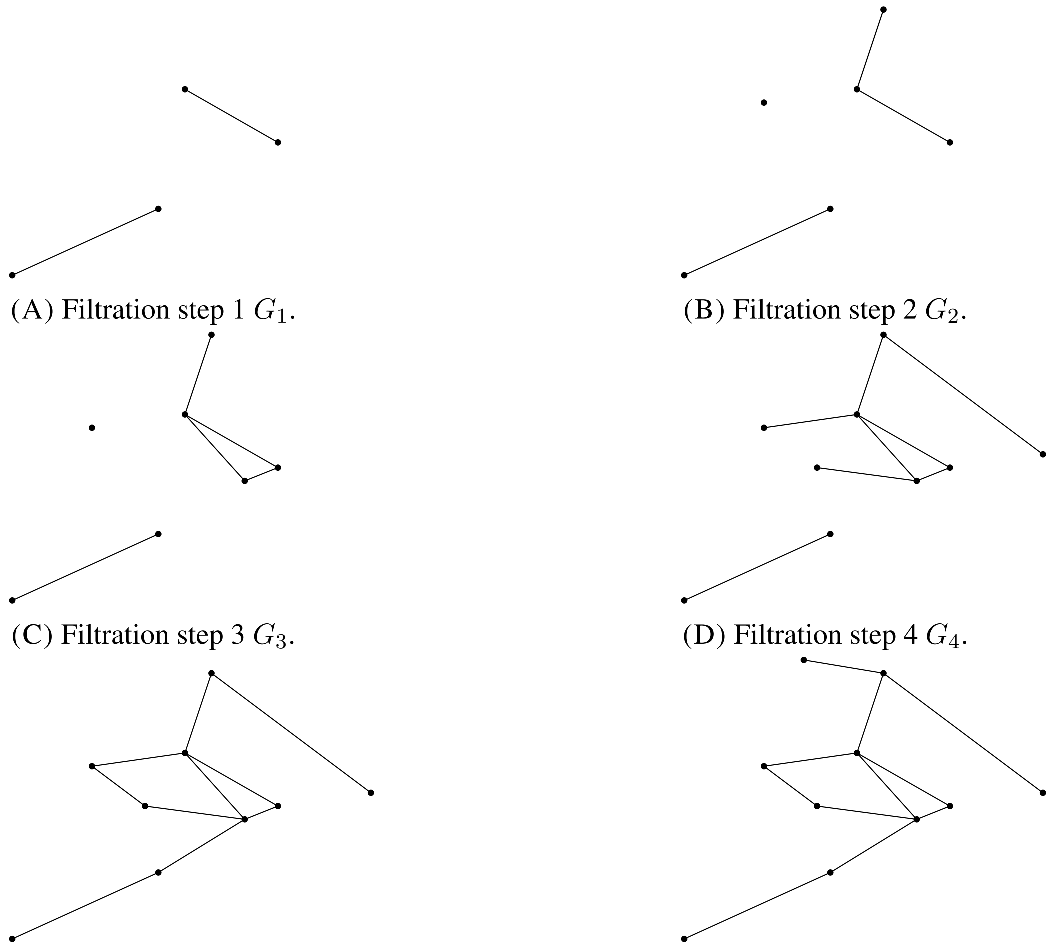
$$\mathfrak{G}^1 = G_0^1 \to G_1^1 \to \dots G_{\alpha_1}^1$$

and

$$\mathfrak{G}^2 = G_0^2 \to G_1^2 \to \dots G_{\alpha_2}^2,$$

with persistence diagrams $D_1 = \mathfrak{D}(\mathfrak{G}^1)$, $D_2 = \mathfrak{D}(\mathfrak{G}^2)$, we can use a distance between D_1 and D_2 as a measure for the (dis)similarity between $\mathfrak{G}^1$ and $\mathfrak{G}^2$. We use the *Wasserstein distance* as a metric on persistence diagrams. In the definition of Wasserstein distance, we consider *matchings* between persistence diagrams. Hence, we first define a matching.

Definition 3.1.5. *Let D_1, D_2 be persistence diagrams. Then a* matching γ *between D_1, D_2 is a bijective map $\Gamma : D_1 \cup \Lambda \to D_2 \cup \Lambda$, with Λ infinitely many copies of the diagonal points (x, x), $x \in \mathbb{R}_{\geq 0}$, as a reservoir of null matches to make Γ bijective.*



(A) Filtration step 1 G_1.

(B) Filtration step 2 G_2.

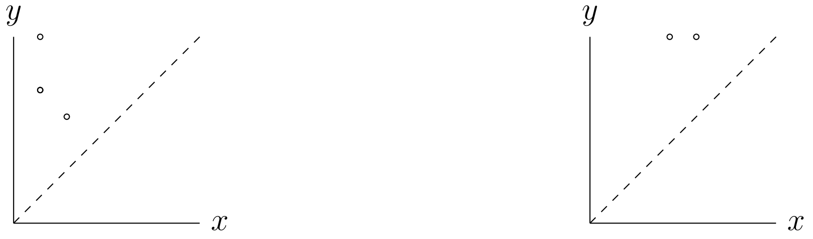
(C) Filtration step 3 G_3.

(D) Filtration step 4 G_4.

(E) Filtration step 5 G_5.

(F) Filtration step 6 G_6.

FIGURE 3.3. Example of a filtration on G.



(A) $\mathcal{D}_0(\mathfrak{G})$.

(B) $\mathcal{D}_1(\mathfrak{G})$.

FIGURE 3.4. $\mathcal{D}(\mathfrak{G})$.

Remark 3.1.6. As matchings are bijective, whether we consider a function from D_1 to D_2 or D_2 to D_1 does not matter.

Definition 3.1.7 (Wasserstein Distance)**.** *Let $X, Y \subset \mathbb{R}^2$ be two multi-sets of points. The 2-Wasserstein distance between X and Y is*

$$\mathcal{W}_2(X,Y) = \left(\inf_{\Gamma \in \mathfrak{M}} \sum_{x \in X} \|x - \Gamma(x)\|^2 \right)^{1/2},$$

where $\mathfrak{M}$ is the set of matchings from X to Y.

3.2. Analysis

The data we have at hand consists of three sets of cells from the same donor, purchased from Lonza (catalogue #: PT-2501). Each set was cultured on glass for 24 hours, after which they were fixed, stained and imaged. We treat these as four data sets in total, one for each set of cells, and then a fourth combining the three experiments. The same analysis was performed on all four sets.

Consider a set X of primary hMSCs from the same donor cultured in the same experimental conditions. For a single cell, let G be the graph representing its boundary. Further, let f be the radial function that returns the distance to the centre of the cell. The persistence diagram of the sub-level sets of f restricted to G provides a summary of the morphological features of the cell.

Remark 3.2.1. We took the centre of the cell to be the centre of mass of the nucleus.

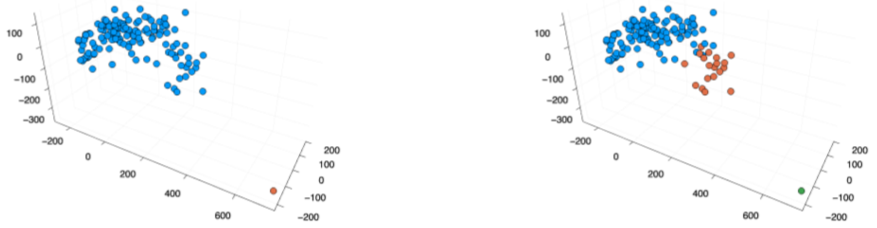
Using the 2-Wasserstein distance, we construct a matrix M of pairwise distances between each pair of cells x_1 and x_2 in X. Then we use standard hierarchical clustering with average linkage to cluster the cells.

We do this with the three sets X_1, X_2, X_3.

3.2.1. Data set X_1.

We begin by looking at the cells in X_1, and seeing if there are any distinct outliers with respect to the Wasserstein metric.

To validate the outlier identified when $k = 2$, we cluster again with $k = 3$, and see the same cell, part1-031, is in its own cluster. We remove part1-031, to obtain X_1' and then perform the clustering again, with $k = 2$.



(A) With $k = 2$, embedded using multidimensional scaling. There is a clear outlier with respect to the Wasserstein distance between the persistence diagrams.

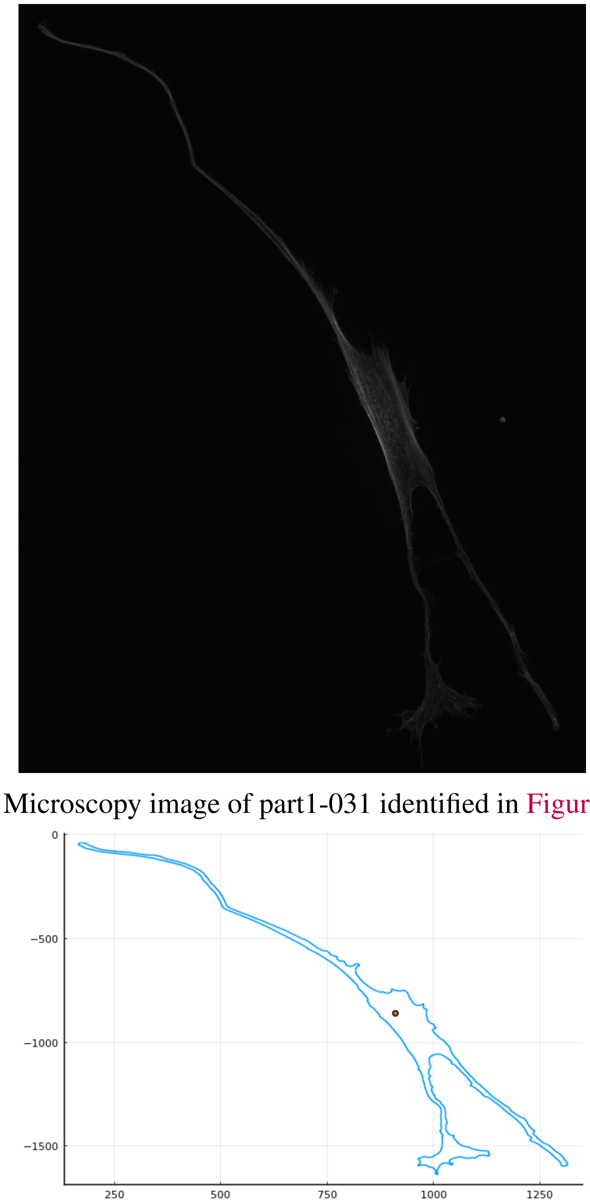
(B) With $k = 3$, embedded using multidimensional scaling. There is a clear outlier with respect to the Wasserstein distance between the persistence diagrams.

FIGURE 3.5. Clustering of cells in X_1'.

There are 139 cells in X_1', with 132 in the main cluster, and 7 in the second cluster, giving a sub-population of 5.035%. Examples of cells in the sub-population are

1. part1-015, Figure 3.9,
2. part1-019, Figure 3.10,
3. part1-023, Figure 3.11,
4. part1-105, Figure 3.12.

We include images (microscopy and contour) of four of the cells in the subpopulation, and then four contours of cells in the main group, see Figure 3.13.



(A) Microscopy image of part1-031 identified in Figure 3.5.

(B) Plot of the boundary part1-015 with the centre marked.

FIGURE 3.6. ‘The protrusions are abnormally long and additionally the kink points towards a potential fixation problem.’ F. Rehfeldt

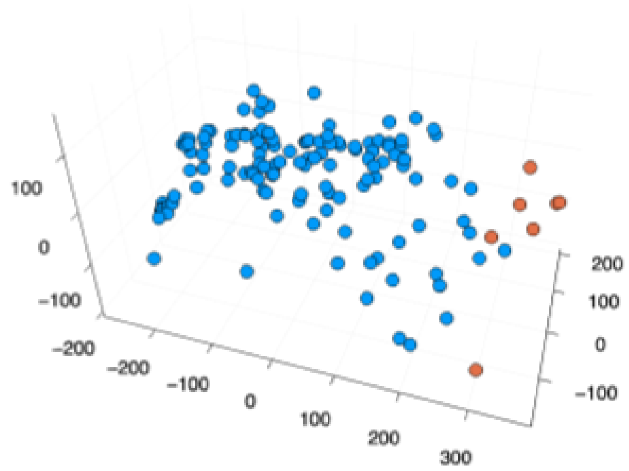
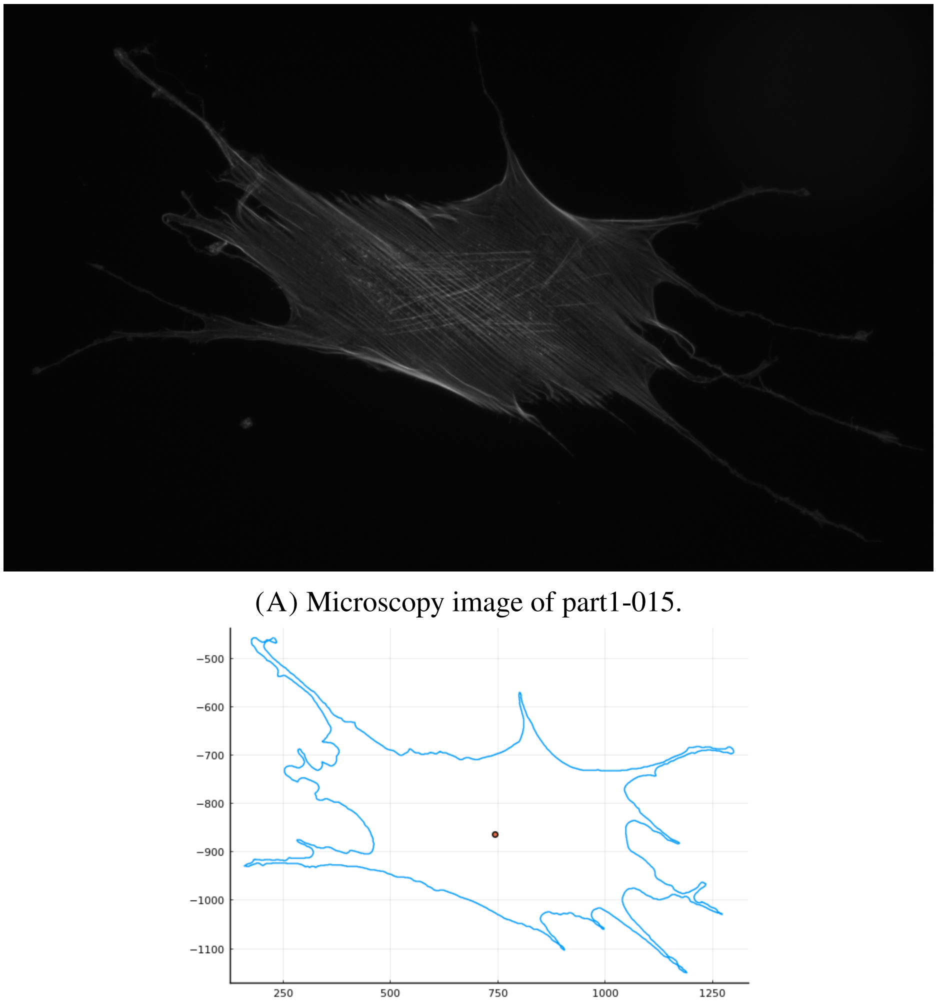


FIGURE 3.7. Clustering of cells in X_1' with $k = 2$, embedded using multidimensional scaling.

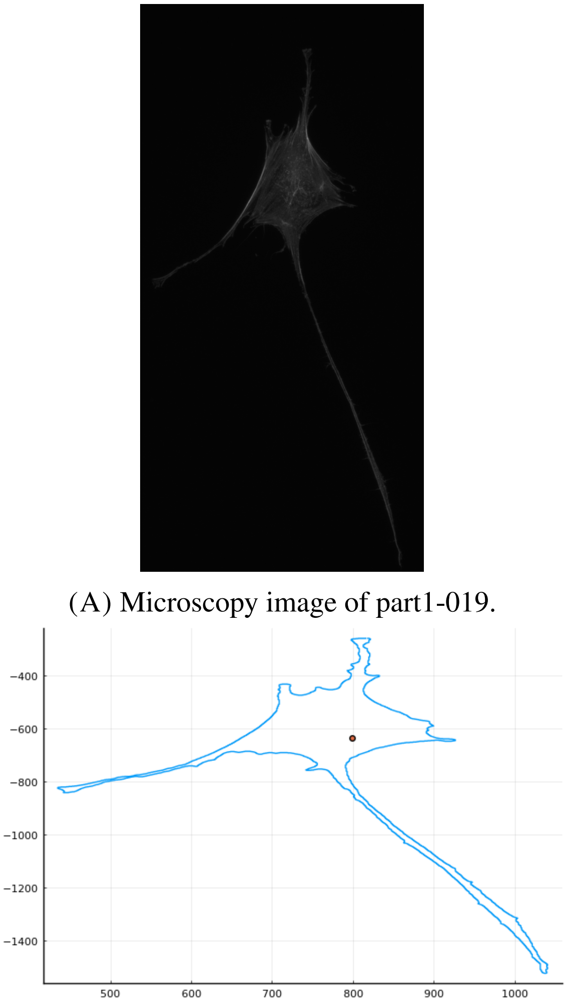
FIGURE 3.8



(A) Microscopy image of part1-015.

(B) Plot of the boundary part1-015 with the centre marked.

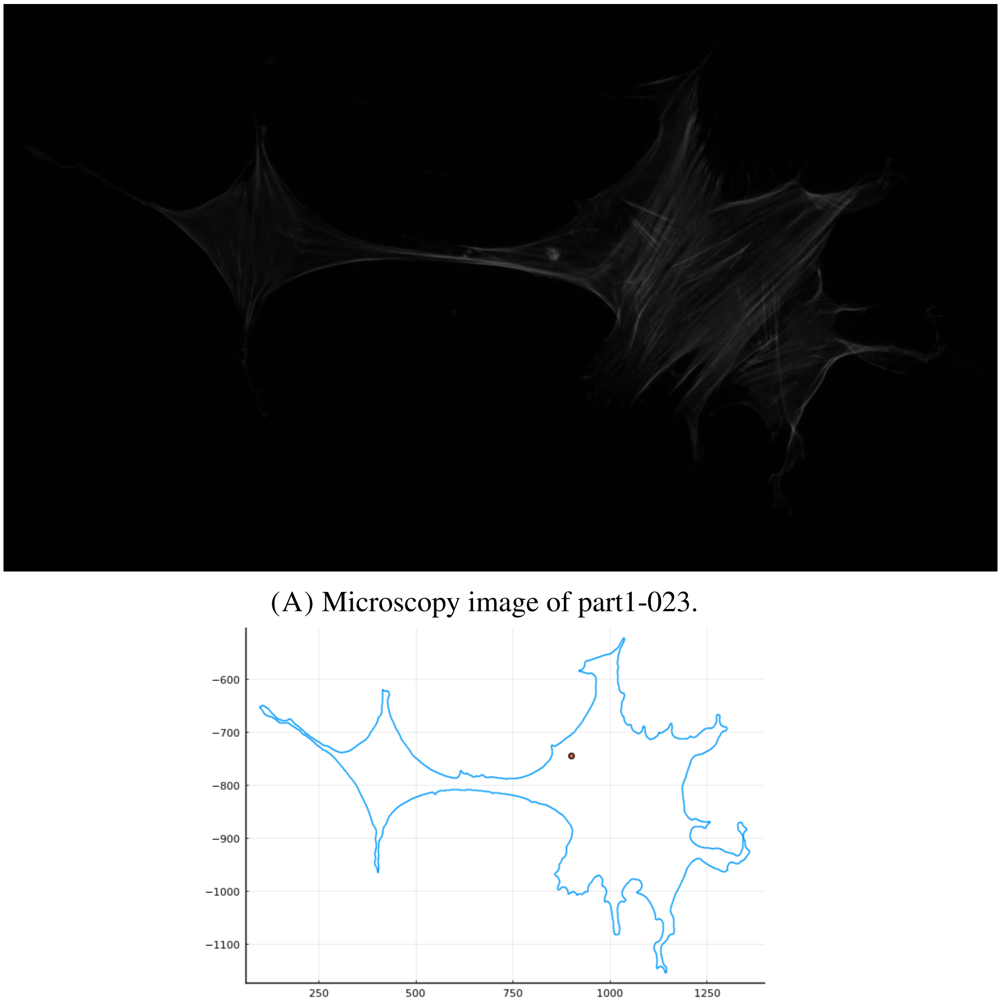
FIGURE 3.9. ‘Abnormal some filaments seem loose (curved) and not properly attached.’ F. Rehfeldt



(A) Microscopy image of part1-019.

(B) Plot of the boundary part1-019 with the centre marked.

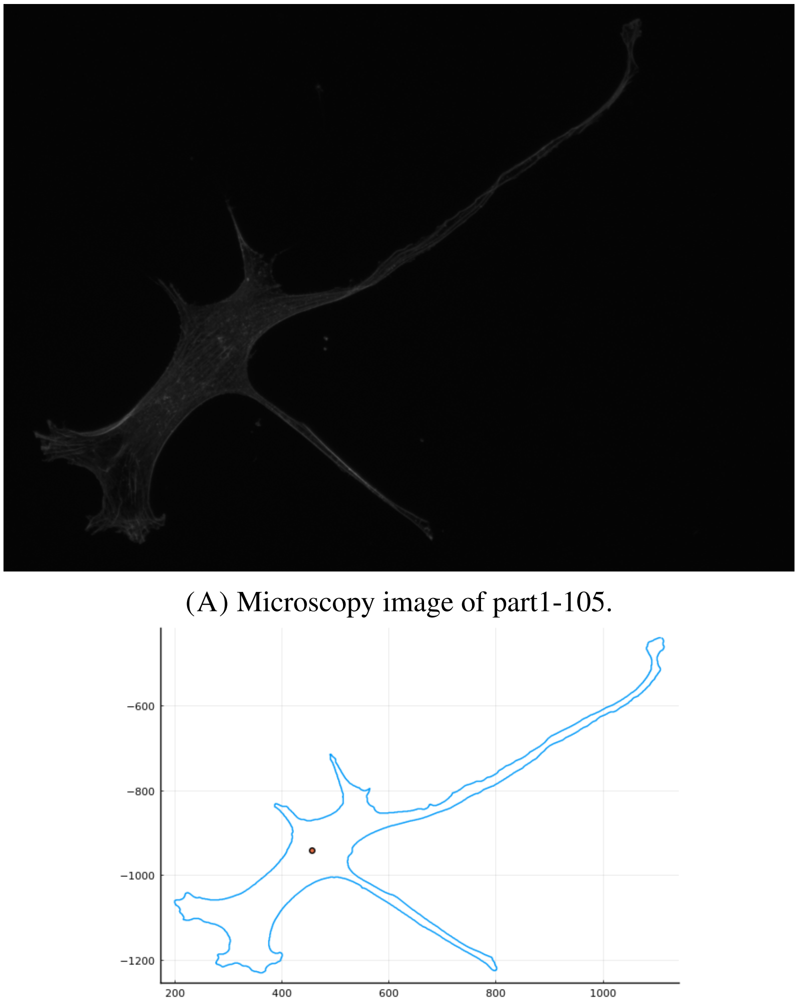
FIGURE 3.10. ‘Cell itself o.k but extremely long protrusions (probably pinned at some spot).’ F. Rehfeldt



(A) Microscopy image of part1-023.

(B) Plot of the boundary of part1-023 with the centre marked.

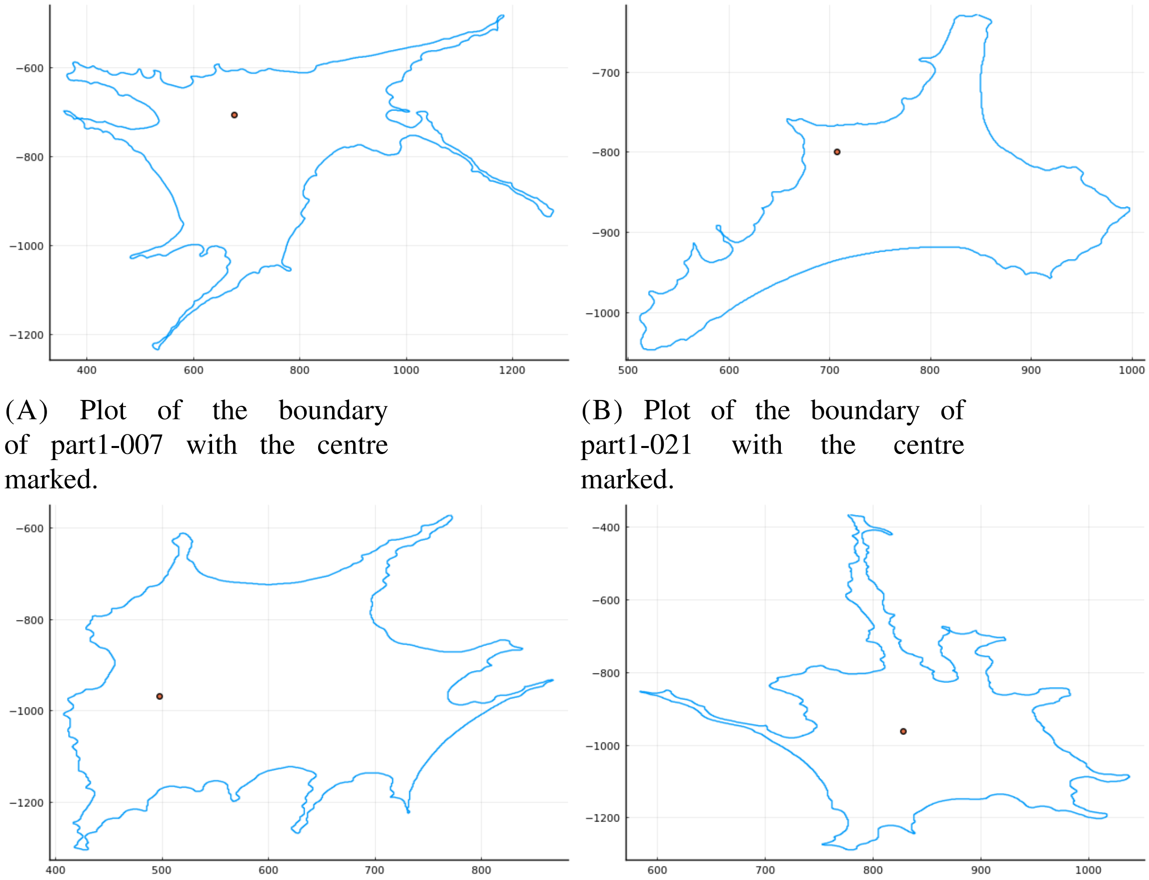
FIGURE 3.11. 'Abnormal cell (two bodies) and some fixation problems (loose filaments).' F. Rehfeldt



(A) Microscopy image of part1-105.

(B) Plot of the boundary of part1-105 with the centre marked.

FIGURE 3.12. ‘Abnormal cell slightly curved/bent long filament that is probably not properly adhered.’ F. Rehfeldt



(A) Plot of the boundary of part1-007 with the centre marked.

(B) Plot of the boundary of part1-021 with the centre marked.

(C) Plot of the boundary of part1-047 with the centre marked.

(D) Plot of the boundary of part1-101 with the centre marked.

FIGURE 3.13. Examples of normal cells in X_1'.

3.2.2. Data set X_2. We begin by looking at the cells in X_2, and seeing if there are any distinct outliers with respect to the Wasserstein metric. Looking at Figure 3.14, there are no cells that appear to be outliers, and so we take these two clusters as our main population and subpopulations. There are 117 cells in X_2, with 15 of these in the secondary cluster, giving a subpopulation of 12.82%.

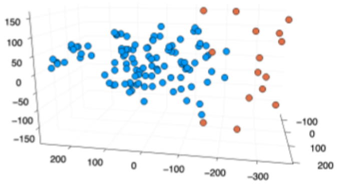
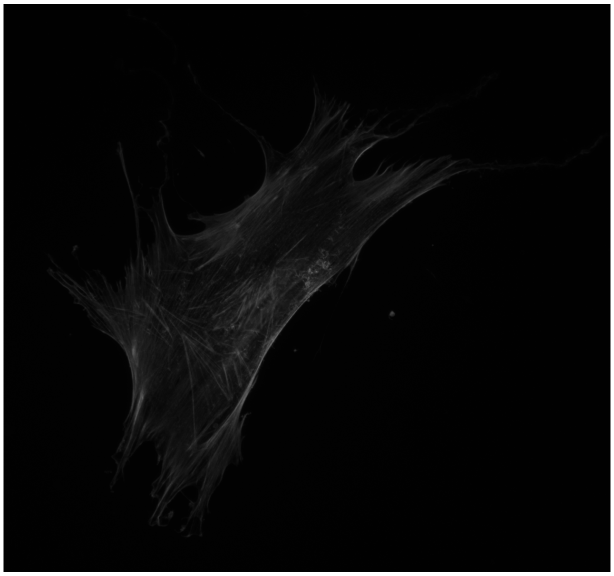


FIGURE 3.14. Clustering of cells in X_2 with $k = 2$, embedded using multidimensional scaling. There are no clear outliers with respect to the Wasserstein distance between the persistence diagrams.

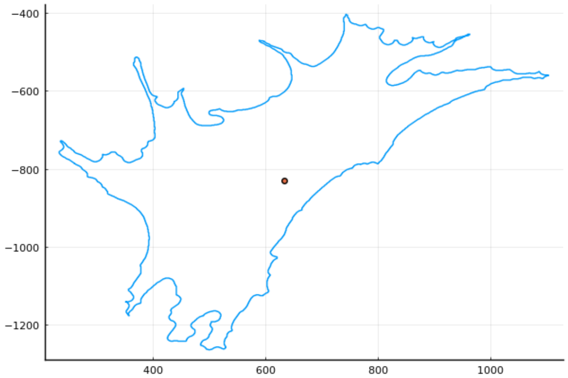
Examples of cells in the subpopulation are:

- part2-067, Figure 3.15,
- part2-073, Figure 3.16,
- part2-097, Figure 3.17,
- part2-213, Figure 3.18.

We include images (microscopy and contour) of four of the cells in the subpopulation, and then four contours of cells in the main group, see Figure 3.19.

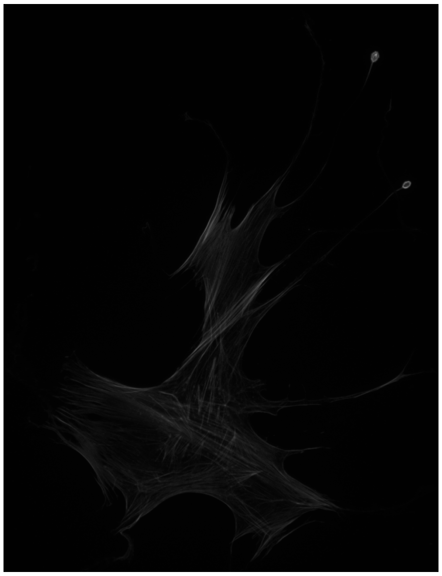


(A) Microscopy image of part2-067.

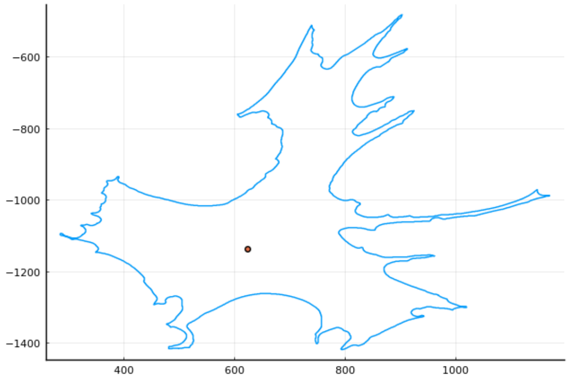


(B) Plot of the boundary of part2-067 with the centre marked.

FIGURE 3.15. ‘Cell is o.k. but also clearly fixation problems are visible.’ F. Rehfeldt

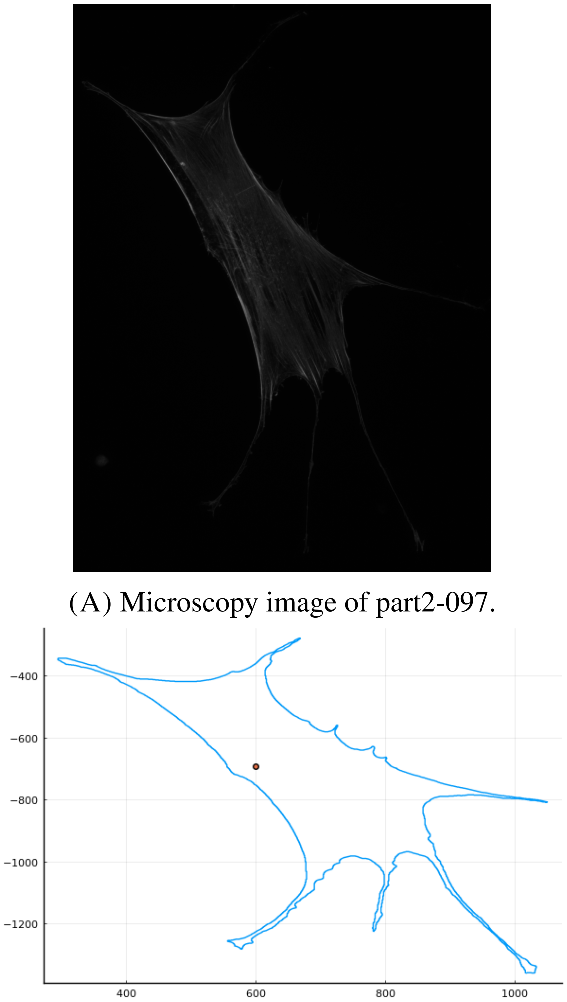


(A) Microscopy image of part2-073.



(B) Plot of the boundary of part2-073 with the centre marked.

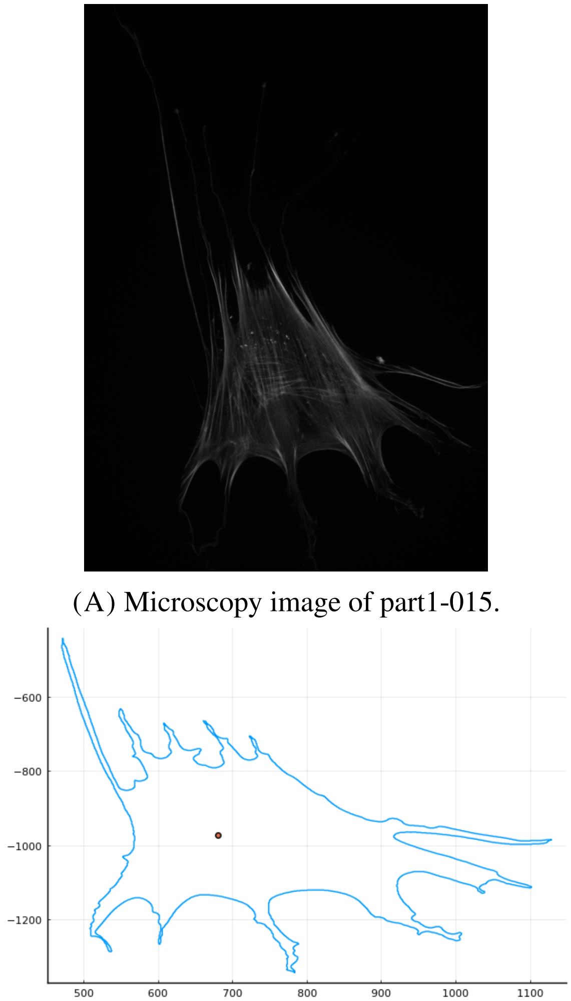
FIGURE 3.16. ‘Clearly abnormal and also some fixation problems.’ F. Rehfeldt



(A) Microscopy image of part2-097.

(B) Plot of the boundary part2-097 with the centre marked.

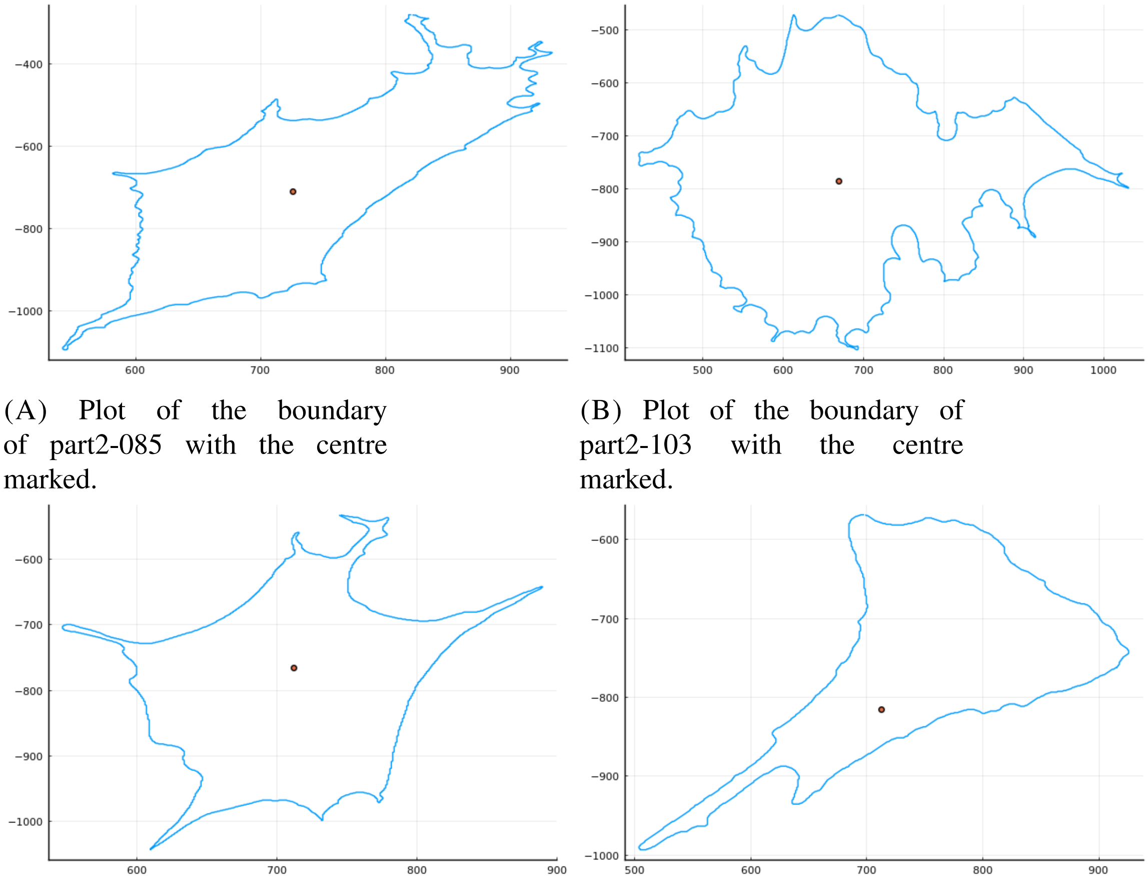
FIGURE 3.17. ‘Cell looks o.k. but as several long protrusions.’ F. Rehfeldt



(A) Microscopy image of part1-015.

(B) Plot of the boundary part1-015 with the centre marked.

FIGURE 3.18. 'Kind of abnormal, fixation problems (bent filaments) and very long protrusions.' F. Rehfeldt



(A) Plot of the boundary of part2-085 with the centre marked.

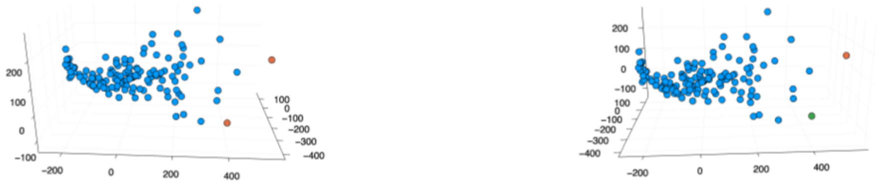
(B) Plot of the boundary of part2-103 with the centre marked.

(C) Plot of the boundary of part2-139 with the centre marked.

(D) Plot of the boundary of part2-219 with the centre marked.

FIGURE 3.19. Examples of normal cells in X_2.

3.2.3. Data set X_3. We begin by looking at the cells in X_3, and seeing if there are any distinct outliers with respect to the Wasserstein metric. There are two outliers identified in Figure 3.1, which we remove to obtain the set X_3'.



(A) Clustering of cells in X_3 with $k = 2$, embedded using multidimensional scaling. There are two potential outliers.

(B) Clustering of cells in X_3 with $k = 2$, embedded using multidimensional scaling. There are two outliers

FIGURE 3.20. Clustering of cells in X_3 with $k = 2$ and $k = 3$, validating the presence of two outliers.

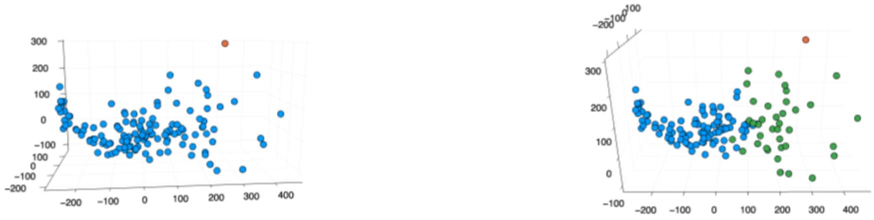
We repeat the same process, Figure 3.21 and identify a further outlier, which we remove to obtain X_3''.

The same process on X_3'' with $k = 2$, produces no potential outliers, and a clear subpopulation consisting of 14 cells, out of the remaining 137, giving a sub-population of 10.22%.

Examples of the cells in the sub-population are

1. part3-051, Figure 3.23,
2. part3-127, Figure 3.24,
3. part3-219, Figure 3.25,
4. part3-231, Figure 3.26.

We include images (microscopy and contour) of four of the cells in the subpopulation, and then four contours of cells in the main group, see Figure 3.27.



(A) Clustering of cells in X_3' with $k = 2$, embedded using multidimensional scaling. There is a potential outlier.

(B) Clustering of cells in X_3' with $k = 2$, embedded using multidimensional scaling. There are two outliers

FIGURE 3.21. Clustering of cells in X_3' with $k = 2$ and $k = 3$, validating the presence of an outlier.

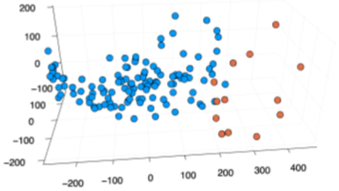
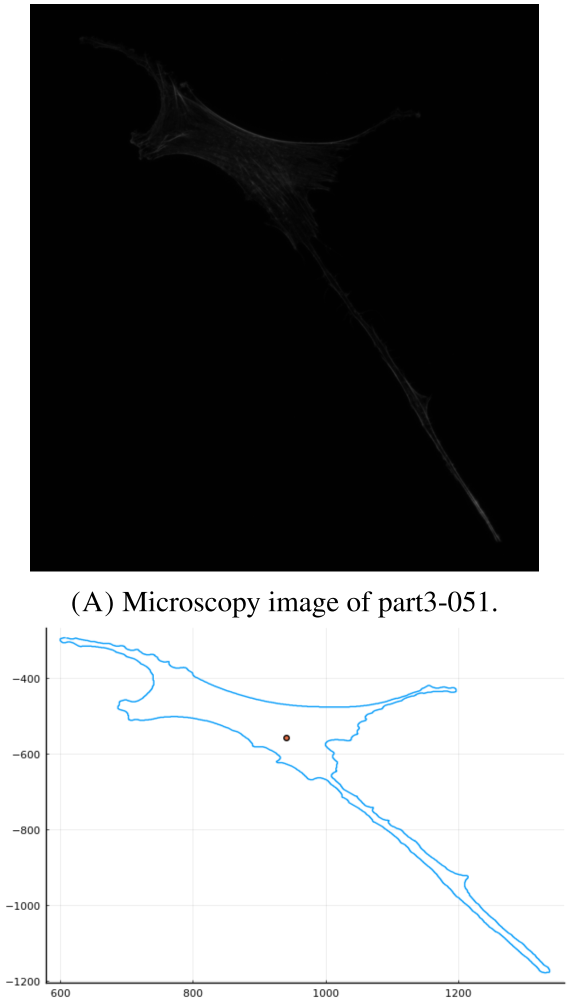


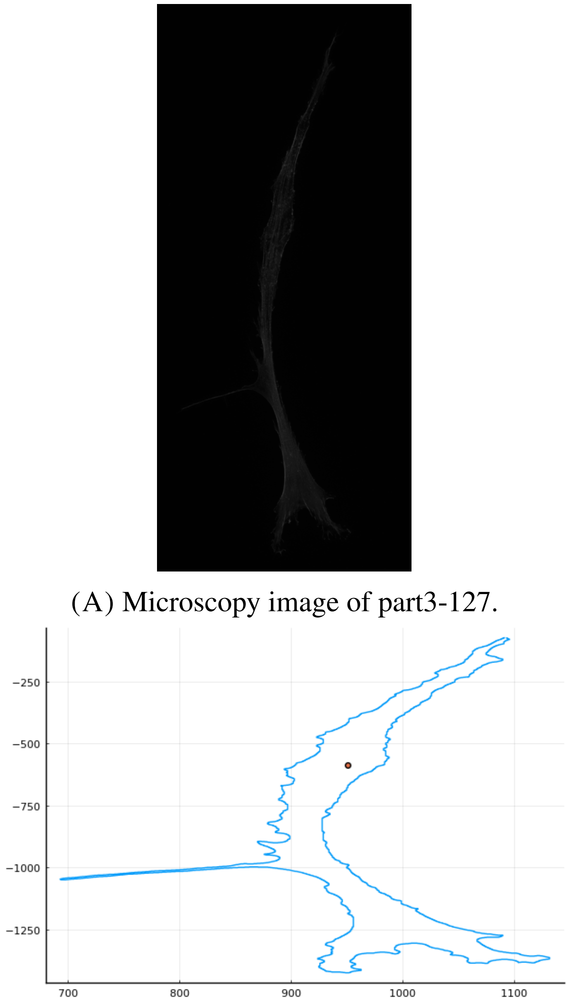
FIGURE 3.22. Clustering of cells in X_3'' with $k = 2$, embedded using multidimensional scaling.



(A) Microscopy image of part3-051.

(B) Plot of the boundary of part3-051 with the centre marked.

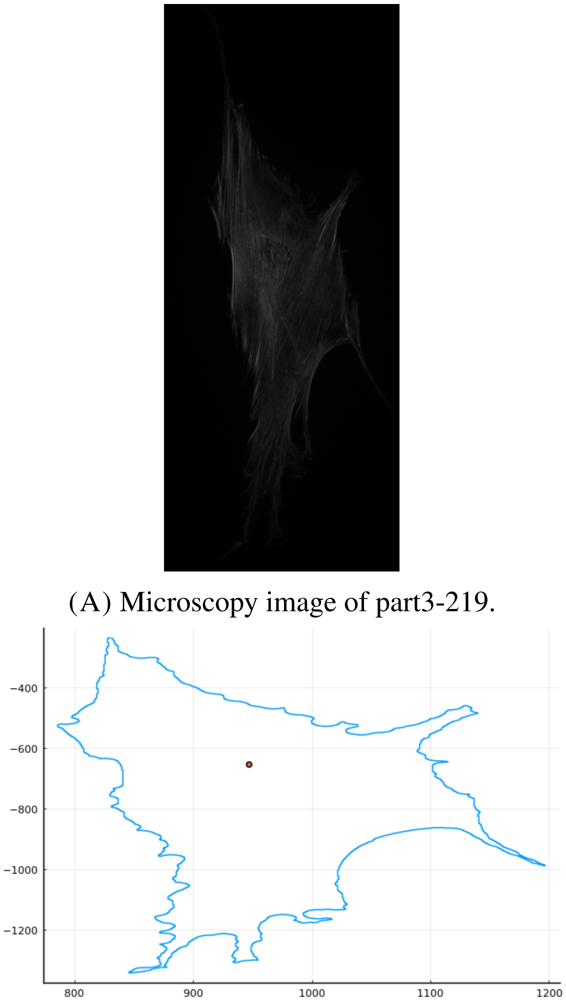
FIGURE 3.23. ‘Extremely elongated and thin protrusion.’ F. Rehfeldt



(A) Microscopy image of part3-127.

(B) Plot of the boundary of part3-127 with the centre marked.

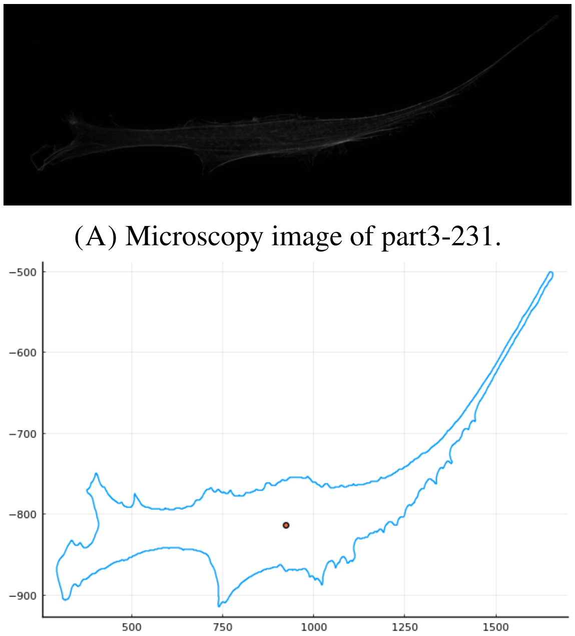
FIGURE 3.24. ‘Extremely elongated and thin, with a very thin protrusion.’ F. Rehfeldt



(A) Microscopy image of part3-219.

(B) Plot of the boundary of part3-219 with the centre marked.

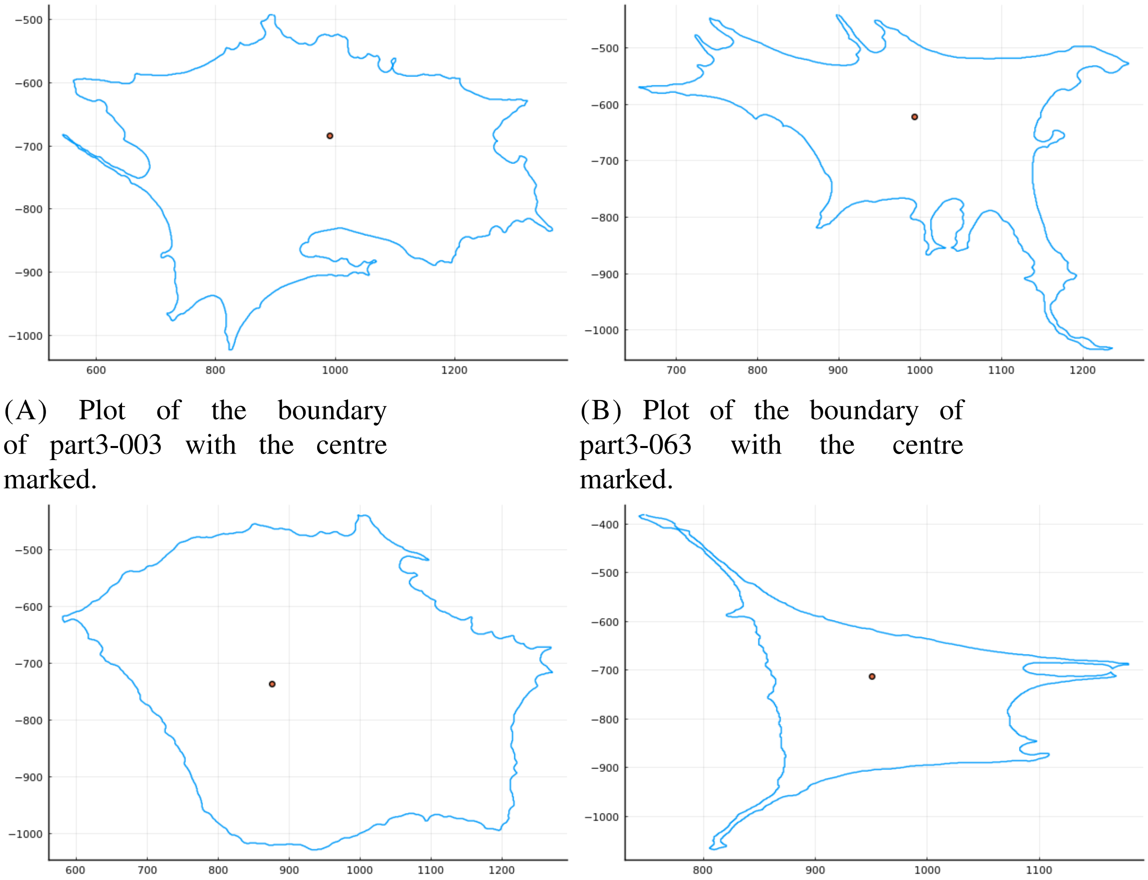
FIGURE 3.25. ‘Has two thin and long protrusions, kinked.’ F. Rehfeldt



(A) Microscopy image of part3-231.

(B) Plot of the boundary of part3-231 with the centre marked.

FIGURE 3.26. 'Extremely elongated, and potential fixation issue in the tail.' F. Rehfeldt



(A) Plot of the boundary of part3-003 with the centre marked.

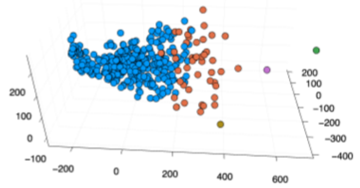
(B) Plot of the boundary of part3-063 with the centre marked.

(C) Plot of the boundary of part3-181 with the centre marked.

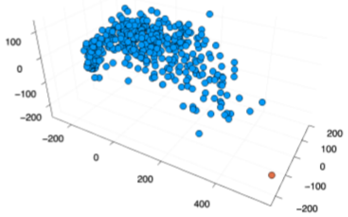
(D) Plot of the boundary of part3-221 with the centre marked.

FIGURE 3.27. Examples of normal cells in X_3.

3.2.4. Combined data set. We now combine the cells in X_1, X_2, X_3 into a single data set, X_4, and compare the results with our analysis on the sets individually. There are some outliers identified in Section 3.2.4, which we remove to obtain the set X_4'. Again, there is another outlier in X_4', which we remove obtaining X_4''.



(A) Clustering of cells in X_4 with $k = 5$, embedded using multidimensional scaling.



(B) Clustering of cells in X_4' with $k = 2$, embedded using multidimensional scaling. There are two outliers

We then examine the clusters in X_4'' with $k = 2$. There are 47 cells in the sub-population, with X_4'' consisting of 392 cells, giving a sub-population of 11.99%.

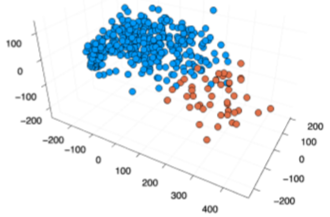


FIGURE 3.29. Clustering of cells in X_4'' with $k = 2$.

Comparing the sub-populations identified in X_1, X_2, X_3 and in the combined X_4, there is a high overlap. It is to be expected that a small number of cells will be in the sub-population in X_4 but not in X_1, X_2, X_3 respectively, as well as the other way. The consistency between the individual and combined analysis indicates that each of the individual data sets is valid.

3.3. Conclusion

In this chapter, we have presented a method for examining the morphological structure of the (2-dimensional) boundary of a cell, to identify 1) any potential outliers, and 2) identify subpopulations. We obtain a summary of the morphology by taking the persistent homology of the boundary, using a radial function from the 'centre' of the cell. We then use the 2-Wasserstein metric on resulting persistence diagrams as a measure of (dis)similarity. To identify outliers and subpopulations, we use average linkage hierarchical clustering and used multi-dimensional scaling to visualise the clustering.

To validate the outliers and subpopulations, we use a 'learned expert', who looks at the microscopy images and contour plots of the cells. There is debate about the use of a biological marker to stain cells that are not mesenchymal stem cells. It is future work to compare the cells identified by this staining [18], with those identified by our process.

We applied this method to three sets of cells cultured separately under the same conditions, and then on the combined set. The results remained stable between the individual and combined analyses, indicating that each of the three sets is valid. The size of the four sub-populations indicates that on top of the 5% of cells which are other primary cells, there is another $5-7\%$ of cells that exhibit abnormal growth due to issues in the culturing and fixing process. Next steps include repeating our analysis with sets of cells cultured under other environmental conditions, as well as with cells from other donors.

References

[1] M. Aanjaneya, F. Chazal, D. Chen, M. Glisse, L. Guibas and D. Morozov, 'Metric Graph Reconstruction from Noisy Data', *International Journal of Computational Geometry & Applications* (22), 305-325, 2012.

[2] P. Bendich, B. Wang and S. Mukherjee, 'Towards Stratification Learning through Homology Inference', *AAAI Fall Symposium on Manifold Learning and its Applications (AAAI)*, 2010.

[3] P. Bendich, B. Wang and S. Mukherjee, 'Local Homology Transfer and Stratification Learning', *Proceedings 23rd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2012.

[4] Y. Bokor, D. Grixti-Cheng, M. Hegland, S. Roberts, and K. Turner, 'Stratified Space Learning: Reconstructing Embedded Graphs' *MODSIM2019, 23rd International Congress on Modelling and Simulation*, 69-75, 2019.

[5] Y. Bokor, K. Turner, and C. Williams, 'Reconstructing linearly embedded graphs: A first step to stratified space learning', *Foundations of Data Science*, 2021.

[6] P. Breiding, S. Kališnik, B. Sturmfels, and M. Weinstein, 'Learning algebraic varieties from samples', *Revista Matemática Complutense* (31), 545 - 593, 2018.

[7] S. Brielle, D. Bavli, A. Motzik, Y. Kan-Tor, X. Sun, C. Kozulin, B. Avni, O. Ram and A. Buxboim, 'Delineating the heterogeneity of matrix-directed differentiation toward soft and stiff tissue lineages via single-cell profiling', *Proceedings of the National Academy of Sciences* (118), 2021 doi:10.1073/pnas.2016322118.

[8] G.Carlsson, 'Topological pattern recognition for point cloud data', *Acta Numerica*, 2014.

[9] F. Chazal, V. de Silva, M. Glisse, and S. Oudot, (2016a). 'The Structure and Stability of Persistence Modules', *SpringerBriefs in Mathematics*, Springer, 2016.

[10] S. Cheng,T. Dey and E. Ramos, 'Manifold Reconstruction from Point Samples', *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 1018-1027, 2005.

[11] L. Costa, N. Eiro, M. Fraile, L. Gonzalez, LuisJ. SaÃ¡, P. Garcia-Portabella, B. Vega, J. Schneider, and F. Vizoso, 'Functional heterogeneity of mesenchymal stem cells from natural niches to culture conditions: implications for further clinical uses', *Cellular and Molecular Life Sciences* (2), 44-46, (2021, doi:10.1007/s00018-020-03600-0.

[12] A. Demptersm N. Laird and D. Rubin, 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society: Series B (Methodological)* (39), 1-22, 1977.

[13] T. K. Dey, *Curve and Surface Reconstruction*, 1^{st} Edition, Cambridge University Press, Cambridge , 2007.

[14] T. Dey and Y. Wang, 'Dimension Detection with Local Homology', *Proceedings of the 26th Canadian Conference on Computational Geometry*, 2014.

[15] B. Eltzner, C. Gottschlich, S. Huckemann, F. Rehfeldt, and C. Wollnik, 'A statistical and Biophysical Toolbox to Elucidate Structure and Formation of Stress Fibers', *Nanoscale Photonic Imaging*, Springer, 263-282, 2020.

[16] A. Hatcher, *Algebraic Topology*, Cambridge University Press, Cambridge, 2000.

[17] B Hill, 'The Persistent Homology Transform and Leaf Shape Analysis', Australian National University, 2020.

[18] F.-J. Lv, R. S. Tuan, K. M. C. Cheung and V. Y. L. Leung, 'Concise Review: The Surface Markers and Identity of Human Mesenchymal Stem Cells', *Stem Cells* (32), 1408-1419, 2014, 10.1002/stem.1681.

[19] D. Phinney 'Functional heterogeneity of mesenchymal stem cells: Implications for cell therapy', *Journal of Cellular Biochemistry* (113), 2806-2812, 2012, 10.1002/jcb.24166.

[20] J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Reuden, S. Saalfeld, B. Schmid, J-T. Tinevez, D. White, V. Hartenstein, K. Eliceiri, P. Tomancak, A. Cardona, 'Fiji: an open-source platform for biological-image analysis', doi:10.1038/nmeth.2019, *Nature Methods* (9), 676-682, 2012.

[21] E. Stein and R. Shakarchi, *Real analysis: measure theory, integration, and Hilbert spaces*, Princeton University Press, Princeton (2009).

[22] B. Stolz, J. Tanner, H. Harrington and V. Nanda, 'Geometric anomaly detection in data', *Proceedings of the National Academy of Sciences* **117 (33)**, 19664-19669, 2020.

[23] K. Turner and S. Mukherjee and D. Boyer, 'Persistent homology transform for modeling shapes and surfaces', *Information and Inference*, (3), 310-314, 2014.

[24] K. Turner and V. Robins and J. Morgan, 'The Extended Persistent Homology Transform of manifolds with boundary', *arXiv*, 2208.14583, 2022.

Alphabetical Index