Abstract

A metric graph is a 1-dimensional stratified metric space consisting of vertices and edges or loops glued together. Metric graphs can be naturally used to represent and model data that take the form of noisy filamentary structures, such as street maps, neurons, networks of rivers and galaxies. We consider the statistical problem of reconstructing the topology of a metric graph from a random sample. We derive a lower bound on the minimax risk for the noiseless case and an upper bound for the special case of metric graphs embedded in \( \mathbb{R}^2 \). The upper bound is based on the reconstruction algorithm given in Aanjaneya et al. (2012).

Keywords: Metric Graph, Filament, Reconstruction, Manifold Learning, Minimax Estimation

1. Introduction

We are concerned with the problem of estimating the topology of filamentary data structure. Datasets consisting of points roughly aligned along intersecting or branching filamentary paths embedded in 2 or higher dimensional spaces have become an increasingly common type of data in a variety of scientific areas. For instance, road reconstruction based on GPS traces, localization of earthquakes faults, galaxy reconstruction are all instances of a more general problem of estimating basic topological features of an underlying filamentary structure. The recent paper by Aanjaneya et al. (2012), upon which our work is based, contains further applications, as well as numerous references. To provide a more concrete example, consider Figure 1. The left hand side displays raw data portraying a neuron from the hippocampus of a rat (Gulyás et al. 1999). The data were obtained from NeuroMorpho.Org (Ascoli et al. 2007). The right hand side of the figure shows the output of the metric graph reconstruction obtained using the algorithm analyzed in this paper, originally proposed by Aanjaneya et al. (2012). The reconstruction, which takes the form of a graph, captures perfectly all the topological features of the neuron, namely, the relationship between the edges and vertices, the number of branching points and the degree of each node.

Metric graphs provide the natural geometric framework for representing intersecting filamentary structures. A metric graph embedded in a \( D \)-dimensional Euclidean space \( (D \geq 2) \) is a 1-dimensional stratified metric space. It consists of a finite number of points (0-dimensional strata) and curves (1-dimensional strata) of finite length, where the boundary
of each curve is given by a pair (of not-necessarily distinct) vertices (see the next section for a formal definition of a metric graph).

In this paper we study the problem of reconstructing the topology of metric graphs from possibly noisy data, from a statistical point of view. Specifically, we assume that we have a sample of points from a distribution supported on a metric graph or in a small neighborhood and we are interested in recovering the topology of the corresponding metric graph. To this end, we use a slightly modified version of the metric graph reconstruction algorithm given in Aanjaneya et al. (2012). Furthermore, in our theoretical analysis we characterize explicitly the minimal sample size required for perfect topological reconstruction as a direct function of parameters defining the shape of the metric graph, introduced in Section 2. This leads to an upper bound on the risk of topological reconstruction in dimension $D = 2$, which we conjecture hold also in higher dimension. Finally, we obtain a lower bound on the risk of topological reconstruction, valid in arbitrary dimensions and for data observed exactly along the metric graph. The lower bound almost matches the derived upper bound, indicating that the algorithm of Aanjaneya et al. (2012) behaves nearly optimally.

Outline. In Section 2 we formally define metric graphs, the statistical models we will consider and the assumptions we will use throughout. We will also describe several geometric quantities that are central to our analysis. Section 3 contains detailed analysis of the performance of algorithm of Aanjaneya et al. (2012) for metric graph reconstruction, under modified settings and assumptions. In Section 4 we derive lower and upper bounds for the minimax risk of metric graph reconstruction problem. Two examples are considered in Section 5. The first is map reconstruction and the second is the neuron example mentioned above. In Section 6 we conclude with some final comments.

Related Work. The work most closely related to ours is Aanjaneya et al. (2012) which was, in fact, the motivation for our work. The algorithm we analyze is in fact a minor modification of the metric graph reconstruction algorithm analyzed by those authors. From the theoretical side, we replace the key assumption in Aanjaneya et al. (2012) of the sample being a $(\varepsilon, R)$-approximation to the underlying metric graph, by the milder assumption of the sample being dense in a neighborhood of the metric graph. Metric graph reconstruction is related to the problem of estimating stratified spaces (basically, intersecting manifolds). Stratified spaces have been studied by a number of authors such as Bendich et al. (2010, 2012); Bendich (2008). A spectral method for estimating intersecting structures is given in Arias-Castro et al. (2011). There are a variety of algorithms for specific problems, for example, see Ahmed and Wenk (2012); Chen et al. (2010) for the reconstruction of road networks. Finally, Chernov and Kurlin (2013) derived an alternative algorithm that uses ideas from homology.

2. Background and Assumptions

The assumptions in Aanjaneya et al. (2012) lead to a reconstruction process that is aimed at capturing the intrinsic structure of the data and is somewhat oblivious to its extrinsic embedding. By considering data embedded in the Euclidean space and focusing on the topological aspect, we show that weaker assumptions are needed to guarantee a correct reconstruction.
Statistical Analysis of Metric Graph Reconstruction

Figure 1: Left: A neuron from the hippocampus of a rat. Right: A metric graph reconstruction of the neuron. The data are from NeuroMorpho.Org (Ascoli et al., 2007).

In this section we provide background on metric graph spaces and describe the assumptions and the geometric parameters that we will be using throughout.

Informally, a metric graph is a collection of vertices and edges glued together in some fashion. Here we state the formal definitions of path metric space and metric graph. For more details see Aanjaneya et al. (2012) and Kuchment (2004).

Definition 1 A metric space \((G,d_G)\) is a path metric space if the distance between any pair of points is equal to the infimum of the lengths of the continuous curves joining them. A metric graph is a path metric space \((G,d_G)\) that is homeomorphic to a 1-dimensional stratified space. A vertex of \(G\) is a 0-dimensional stratum of \(G\) and an edge of \(G\) is a 1-dimensional stratum of \(G\).

We will consider metric graphs embedded in \(\mathbb{R}^D\). Note that, if one ignores the metric structure, namely the length of edges and loops, the shape or topology of a metric graph \((G,d_G)\) is encoded by a graph, whose vertices and edges correspond to vertices and edges of \(G\). Since we allow for two vertices to be connected by more than one edge we are actually dealing with pseudographs. We recall that an undirected pseudograph \((V,E)\) is a set of vertices \(V\), a multiset \(E\) of unordered pairs of (not necessarily distinct) vertices. To a given pseudograph we can associate a function \(f : E \to V \times V\), which, when applied to an edge \(e \in E\), simply extracts the vertices to which \(e\) is adjacent. Thus, if \(e_1, e_2 \in E\) are such that \(f(e_1) = f(e_2)\), then \(e_1\) and \(e_2\) are parallel edges. Similarly, if \(e \in E\) is such that \(f(e) = \{v,v\}\) for some \(v \in V\), then \(e\) is a loop. For each pair \((u,v) \in V \times V\), let \(\nu(u,v) = |f^{-1}(\{u,v\})|\) if \(\{u,v\} \in E\) and 0 otherwise. In particular, \(\nu(u,v)\) is the number of edges between \(u\) and \(v\) (or loops if \(u = v\)).

3
We say that a metric graph reconstruction algorithm perfectly recovers the topology of $G$ if outputs a pseudograph isomorphic to the pseudograph representing the topology of $G$.

We now define some key quantities regarding the structure of a metric graph. We start with the general definition of condition number. See \cite{ChazalLieutier2006} and \cite{Niyogi2008}.

**Definition 2** The condition number of a 1-dimensional manifold $M$ embedded in $\mathbb{R}^D$, with boundary $\{v,v'\}$, is the largest number $\tau$ such that the open normal bundle about $M\setminus\{v,v'\}$ of radius $r$ is embedded in $\mathbb{R}^D$ for every $r < \tau$.

The condition number is a measure of the curvature of a manifold. A manifold with large condition number does not come too close to be self-intersecting. For example the condition number of an arc of a circle is equal to its radius. Each edge of a metric graph $(G,d_G)$ can be seen as a 1-dimensional manifold with boundary. Let the local condition number be the minimum of the condition numbers associated to the edges of a metric graph. To control points far away in the graph distance, but close in the embedding space, we define

$$A_b = \{(x,x') \in G : d_G(x,x') \geq \min(b, \tau a)\}$$

and we define the global condition number as the infimum of the Euclidean distances among pairs of point in $A_b$, i.e. $\inf_{A_b} \|x - x'\|$.

When 2 edges intersect at a vertex $v$ they create an angle, where the angle between two intersecting curves is formally defined as follows. Suppose that $e_1$ and $e_2$ intersect at $x$. Let $B(x,\epsilon)$ be the $D$-dimensional ball of radius $\epsilon$ centered at $x$. Let $\ell_1(\epsilon)$ be the line segment joining the two points $x$ and $B(x,\epsilon) \cap e_1$. Let $\ell_2(\epsilon)$ be the line segment joining the two points $x$ and $B(x,\epsilon) \cap e_2$. Let $\alpha_\epsilon(e_1, e_2)$ be the angle between $\ell_1(\epsilon)$ and $\ell_2(\epsilon)$. The angle between $e_1$ and $e_2$ is $\alpha(e_1, e_2) = \lim_{\epsilon \to 0} \alpha_\epsilon(e_1, e_2)$. We assume that, for each pair of intersecting edges $e_1$ and $e_2$, the angle $\alpha(e_1, e_2)$ is well-defined.

Let $(G,d_G)$ be a metric graph and, for a constant $\sigma \geq 0$, let $G \oplus \sigma = \{y : \inf_{x \in G} \|x-y\| \leq \sigma\}$, the $\sigma$ tube around $G$. In particular, if $\sigma = 0$, then, trivially, $G \oplus \sigma = G$. Notice that $G \oplus \sigma$ is a set of dimension $D$ if $\sigma > 0$. The problem of metric graph reconstruction consists of reconstructing a metric graph $G$ given a sample $\{y_1, \ldots, y_n\} \subset \mathbb{Y} \subset G \oplus \sigma$ endowed with a distance $d_\mathbb{Y}$, which could be the $D$-dimensional Euclidean distance or some more complicate distance. If $\sigma = 0$ we say that the sample $\mathbb{Y}$ is noiseless, while if $\sigma > 0$, we say that $\mathbb{Y}$ is a noisy sample.

We will use the assumption that the sample $\mathbb{Y}$ is sufficiently dense in $G \oplus \sigma$ with respect to the Euclidean metric, as formalized below.

**Definition 3** The sample $\mathbb{Y} = \{y_1, \ldots, y_n\} \subset G \oplus \sigma \subset \mathbb{R}^D$ is $\frac{\delta}{2}$-dense in $G \oplus \sigma$ if for every $x \in G \oplus \sigma$, there exists a $y \in \mathbb{Y}$ such that $\|x-y\| < \frac{\delta}{2}$.

Note that this standard assumption is necessary for any recovery guarantee. In this paper we are interested in characterizing how dense a sample needs to be in order to guarantee, with high probability, perfect recovery of the topology of a metric graph. While in our analysis we will mainly rely on the assumption of a dense sample, \cite{Aanjaneya2012} used the more refined but stronger assumption of the sample being an $(\epsilon, R)$-approximation, which we recall.
Definition 4 Given positive numbers $\varepsilon, R$, we say that $(\mathcal{Y}, d_\mathcal{Y})$ is an $(\varepsilon, R)$-approximation of the metric space $(G, d_G)$ if there exists a correspondence $C \subset G \times \mathcal{Y}$ such that

$$(x, y), (x', y') \in C, \min (d_G(x, x'), d_\mathcal{Y}(y, y')) \leq R \implies |d_G(x, x') - d_\mathcal{Y}(y, y')| \leq \varepsilon. \quad (1)$$

As shown in Aanjaneya et al. (2012), the $(\varepsilon, R)$-approximation assumption is sufficient, for appropriate choice of the parameters $\varepsilon$ and $R$, to recover not only the topology of a metric graph $(G, d_G)$, but also its metric $d_G$ with high accuracy. However, when compared to the dense sample assumption, it demands a larger sample complexity to achieve accurate topological reconstruction. We illustrate this claim with a simple example.

We need one more definition. The Rips-Vietoris graph $\mathcal{R}_\kappa(\mathcal{Y})$ is the graph with vertex set $\mathcal{Y}$ and an edge exists between $y_i$ and $y_j$ iff $||y_i - y_j|| \leq \kappa$.

Example 1 Figure 2 shows a metric graph $(G, d_G)$ embedded in $\mathbb{R}^2$ (top left) consisting of three edges of length 1 that intersect at a common vertex $O$ forming 3 angles of width $2\pi/3$ and a $\delta/2$-dense sample $\mathcal{Y} = \{y_1, \ldots, y_n\}$ in $G \oplus 0$ (top right), for some $\delta > 0$.

The authors of Aanjaneya et al. (2012) define $d_\mathcal{Y}(y_1, y_2)$ to be the length of the shortest path connecting $y_1$ and $y_2$ in the Rips-Vietoris graph $\mathcal{R}_\kappa(\mathcal{Y})$. Their key assumption is that $(\mathcal{Y}, d_\mathcal{Y})$ is an $(\varepsilon, R)$-approximation of $(G, d_G)$.

We would like to keep $\kappa$ as small as possible with the constraint $\kappa \geq \delta$, to guarantee that $\mathcal{R}_\kappa(\mathcal{Y})$ is a connected graph. Thus we set $\kappa = \delta$. In this example, the values $\varepsilon = 1/128$ and $R = 5/17$ satisfy Theorems 1 and 2 in Aanjaneya et al. (2012) for topological and metric reconstruction. The reconstruction algorithm, described here in Section 3, provides a reconstructed graph $\hat{G}$ (Figure 3 bottom right) that is isomorphic to $G$ and a distance $d_\hat{G}$ that approximates $d_G$.

Consider the bottom left plot of Figure 3 where $\{x, x'\} \subset G$, $\{y, y', z, z'\} \subset \mathcal{Y}$, $d_\mathcal{Y}(y, y') = R$, $d_\mathcal{Y}(z, z') = \delta$ and $(x, y), (x', y') \in C$. Note that $||O - z|| = \delta/\sqrt{3}$ and

$$|d_G(x, x') - d_\mathcal{Y}(y, y')| = \frac{2\delta}{\sqrt{3}}.$$

Therefore $(\mathcal{Y}, d_\mathcal{Y})$ is a $(1/128, 5/17)$-approximation of the metric graph $(G, d_G)$ if $\delta \leq \frac{\sqrt{3}}{2} \frac{1}{128} = 0.00677$. This strong condition guarantees topological and metric reconstruction. By focusing on the topological aspect, we show that minor modifications of the algorithm guarantee a correct reconstruction with weaker assumptions on the sample $\mathcal{Y}$. In particular, for this example, a 0.08/2-dense sample is sufficient.

Throughout our analysis we restrict the attention to metric graphs embedded in $\mathbb{R}^D$ that satisfy the following assumptions:

A1 The graphs are free of nodes of degree 2 (though they may contain vertices of degree 1 or 3 and higher).

A2 Each edge is a smooth embedded sub-manifold of dimension 1, of length at least $b > 0$ and with condition number at least $\tau > 0$.

A3 Each pair of intersecting edges forms a well-defined angle of size at least $\alpha > 0$. 


Figure 2: Top left: metric graph $\left( G, d_G \right)$ formed by 3 edges of length 1 intersecting at a single vertex and forming 3 angles of width $2\pi/3$. Top right: sample $Y \subset G \oplus 0$. Bottom left: example of the distortion between $d_G(x, x')$ and $d_Y(y, y')$. Bottom right: reconstructed graph $\hat{G}$. 
**A4** The global condition number is at least $\xi > 0$.

Assumptions A1 and A2 allow us to consider each edge of a metric graph as a single smooth curve. This assumption is common in the literature. See Chen et al. (2010) for a road reconstruction algorithm that allows for corners within an edge.

Let $\mathcal{G}$ be the set of metric graphs embedded in $\mathbb{R}^D$ that satisfy assumptions A1, A2, A3 and A4. We consider the collection $\mathcal{P}$ of probability distributions supported over metric graphs $(G, d_G)$ in $\mathcal{G}$ having densities $p$ with respect to the length of $G$ bounded from below by a constant $a > 0$. We are interested in bounding the minimax risk

$$R_n = \inf_{\hat{G}} \sup_{P \in \mathcal{P}} P^n(\hat{G} \not\simeq G),$$

where the infimum is over all estimators $\hat{G}$ of the topology of $(G, d_G)$, the supremum is over the class of distributions $\mathcal{P}$ for $\mathcal{Y}$ and $\hat{G} \not\simeq G$ means that $\hat{G}$ and $G$ are not isomorphic.

In Section 4 we will find a lower bound for $R_n$ and an upper bound for the special case of metric graphs embedded in $\mathbb{R}^2$.

We conclude this section by summarizing the many parameters and symbols involved in our analysis. See Table 1.

| Symbol | Meaning |
|--------|---------|
| $(G, d_G)$ | metric graph |
| $\alpha$ | smallest angle |
| $b$ | shortest edge |
| $\tau$ | local condition number |
| $\xi$ | global condition number |
| $\mathcal{G}$ | set of metric graphs embedded in $\mathbb{R}^D$, satisfying A1-A4 |
| $\mathcal{P}$ | set of distributions on $G$ with density bounded by $a > 0$ |
| $G \oplus \sigma$ | $\sigma$ tube around $G$ |
| $\mathcal{Y}$ | sample, subset of $G \oplus \sigma$ |
| $\delta$ | $\mathcal{Y}$ is a $\delta/2$-dense sample |
| $\mathcal{R}_\kappa(\mathcal{Y})$ | R-V graph on $\mathcal{Y}$ of parameter $\kappa$ |

Table 1: Summary of the symbols used in our analysis.

3. Performance Analysis for the Algorithm of Aanjaneya et al. (2012)

In this section we will study the performance of the metric graph reconstruction algorithm of Aanjaneya et al. (2012), under assumptions A1-A4 and with the choice of parameters adapted to our setting. In Section 4 we will use these results to derive upper bounds on the minimax rate for topology reconstruction for noiseless samples. Our analysis applies to metric graphs embedded in $\mathbb{R}^2$, though we conjecture that it carries over to arbitrary dimensions.
The metric graph reconstruction algorithm is presented in Algorithm 1.

| Algorithm 1: Metric Graph Reconstruction Algorithm |
|-----------------------------------------------------|
| **Input:** sample \( Y, d_Y \). \|
| 1: **Labeling points as edge or vertex** \|
| 2: for all \( y \in Y \) do \|
| 3: \( S_y \leftarrow B(y, r + \delta) \setminus B(y, r) \) \|
| 4: \( \text{deg}_r(y) \leftarrow \text{Number of connected components of Rips-Vietoris graph } R_\delta(S_y) \) \|
| 5: if \( \text{deg}_r(y) = 2 \) then \|
| 6: Label \( y \) as a edge point \|
| 7: else \|
| 8: Label \( y \) as a preliminary vertex point. \|
| 9: end if \|
| 10: end for. \|
| 11: Label all points within Euclidean distance \( p_{11} \) from a preliminary vertex point as vertices. \|
| 12: Let \( E \) be the point of \( Y \) labeled as edge points. \|
| 13: Let \( V \) be the point of \( Y \) labeled as vertices. \|
| 14: **Reconstructing the graph structure** \|
| 15: Compute the connected components of the Rips-Vietoris graphs \( R_\delta(E) \) and \( R_\delta(V) \). \|
| 16: Let the connected components of \( R_\delta(V) \) be the vertices of the reconstructed graph \( \hat{G} \). \|
| 17: Let there be an edge between vertices of \( \hat{G} \) if their corresponding connected components in \( R_\delta(V) \) contain points at distance less than \( \delta \) from the same connected component of \( R_\delta(E) \). \|

The inner radius of the shell at Step 3 and the width of the expansion at Step 11 are parameters the user has to specify. Note that in the original algorithm of Aanjaneya et al. (2012) the expansion of Step 11 was performed using the distance \( d_Y \), since an embedding of the space was not required.

The algorithm takes a (possibly noisy) sample \( Y \) from a metric graph \( G \) and a distance \( d_Y \) defined on \( Y \) and returns a graph \( \hat{G} \) that approximates \( G \). We will analyze the algorithm considering first the Euclidean distance and then specifying a more complex distance \( d_Y \) on the sample \( Y \).

Before providing the details of our analysis, we demonstrate the worst case scenario of two edges \( e_1 \) and \( e_2 \), with minimum condition number \( \tau \) forming the smallest angles \( \alpha \) at vertex \( x \). See Figure 3 (left). Note that \( e_1 \) and \( e_2 \) are simply arcs of circles of radius \( \tau \). \( O \) and \( O' \) are the centers of the circles associated to edges \( e_1 \) and \( e_2 \). It is easy to see that angle \( O\hat{x}O' \) has width \( \pi - \alpha \).

Let \( Y \) be a noisy sample of \( G \). In other words \( Y \) is a subset of \( G \oplus \sigma \), the tube of radius \( \sigma \geq 0 \) around the metric graph \( G \). See Figure 3 (right).

For \( \sigma \geq 0 \), the smallest angle formed by the inner faces of the tube around the metric graph is

\[
\alpha' = \pi - \arccos \frac{2(\tau - \sigma)^2 - 4\tau^2 \cos^2(\alpha/2)}{2(\tau - \sigma)^2},
\]

where we applied the cosine law to the triangle \( OsO' \) and the fact that angle \( O\hat{x}O' \) has width \( \pi - \alpha' \). Note that if \( \sigma = 0 \) then \( \alpha' = \alpha \).
Figure 3: Left: edges $e_1$ and $e_2$ with minimum condition number $\tau$ forming the smallest angles $\alpha$ at vertex $x$. Right: same metric graph with a tube of radius $\sigma$ around it.

Let $Q$ be the midpoint of segment $\overline{OO'}$ and let $T$ be the intersection point of $\overline{OO'}$ and edge $e_1$. It can be shown that

$$x\hat{O}O' = \alpha/2,$$  \hspace{1cm} (3)

$$T\hat{x}Q = \alpha/4.$$  \hspace{1cm} (4)

**Analysis of Algorithm 1 with Euclidean distances**

Consider the framework introduced in Section 2 for metric graphs embedded in $\mathbb{R}^2$ and Algorithm 1 with input $(Y, d_Y = \| \cdot \|)$, that is, the Euclidean distance is used at every step. The algorithm requires the specification of $r$, the inner radius of the shell, and $p_{11}$, the parameter governing the expansion of Step 11. We set

$$r = \max \left\{ \frac{\delta}{2} + \tau \sin(\alpha/2) - (\tau - \sigma) \sin(\alpha'/2) + \frac{\delta}{2 \sin(\alpha/4)}, \right.$$  

$$\sqrt{\tau^2 + (\tau - \sigma)^2 - 2\tau(\tau - \sigma) \sqrt{1 - \left(\frac{\delta}{2(\tau - \sigma)}\right)^2}} \right\}$$
and

$$p_{11} = \begin{cases} \frac{3\delta}{2} + \tau \sin(\alpha/2) - (\tau - \sigma) \sin(\alpha'/2) + r & \text{if } \alpha' \geq \pi/2 \text{ and } \\
\delta + r \leq (\tau - \sigma) \frac{\sin(2\alpha' - \pi)}{\sin(\alpha')} & \text{otherwise.} \end{cases}$$

This choice is justified in the proof of Proposition 5.

**Proposition 5** If \( \mathbb{Y} \) is \( \frac{\delta}{2} \)-dense in \( G \oplus \sigma \) and

\[
0 \leq \sigma < \tau[1 - \cos(\alpha/2)],
\]

\[
0 < r + \delta < \xi - 2\sigma,
\]

\[
\sin\left(\frac{\min(b, \alpha \tau) - (\alpha - \alpha')\tau}{2\tau}\right) > \begin{cases} \frac{r + p_{11} + 3\delta/2}{\tau - \sigma} & \text{if } \alpha' \geq \pi/2 \text{ and } \\
\frac{\delta + p_{11} + \delta/2}{\tau - \sigma} & \text{otherwise}\end{cases}
\]

then the graph \( \hat{G} \) provided by Algorithm 1 (input: \( \mathbb{Y}, \| \cdot \| \)) is isomorphic to \( G \).

**Proof** First, note that condition 5 is a bound on the radius of the tubular noise around the metric graph. Consider the case depicted in Figure 3 (right). The radius of the tube must be less than half of the maximum distance between the represented edges \( \|Q - T\| = \|O' - T\| - \|O' - Q\| = \tau - \tau \cos(\alpha/2) \). Next, condition 6 guarantees that points of \( G \) which are far apart in the metric graph distance \( d_G \) and close in the Euclidean distance do not interfere in the construction of the shells at Steps 3-4.

The rest of the proof involves condition 7. Since the sample is \( \frac{\delta}{2} \)-dense in the tube, there is at least a point \( y \in \mathbb{Y} \) inside the ball of radius \( \frac{\delta}{2} \) centered at any vertex \( x \in G \). When we apply Algorithm 1 we want to be sure that \( y \) is labeled as a vertex, i.e. the number of connected components of the shell around \( y \) is different than 2 (Steps 3-4). The worst case is depicted in Figure 3 (left), where \( x \) is the vertex of the minimum angle \( \alpha \), formed by two edges, \( e_1 \) and \( e_2 \).

The inner faces of the tube of radius \( \sigma \) around \( e_1 \) and \( e_2 \) form an angle of width \( \alpha' \) at vertex \( s \), as described at the beginning of this section. Let \( u \) and \( v \) be the two points on the faces of the tube such that they are equidistant from \( x \) and \( \|u - v\| = \delta \). Since at Step 4 we construct a \( R_{\delta}(S_g) \) graph to determine the number of connected components of the shell \( S_g \) and we want \( y \) to be a vertex, we choose \( r \), the inner radius of the shell, so that if \( u, v \in \mathbb{Y} \) then \( r \geq \max\{d_\gamma(y, u), d_\gamma(y, v)\} \). This guarantees that \( \forall t_1, t_2 \in \mathbb{Y} \) with \( t_1 \) around edge \( e_1 \), \( t_2 \) around edge \( e_2 \) with \( d_\gamma(y, t_1) \geq r \), \( d_\gamma(y, t_2) \geq r \), we have \( d_\gamma(t_1, t_2) \geq \delta \), i.e. \( t_1 \) and \( t_2 \) belong to different connected component of the shell around \( y \) at Step 4.

We bound the distance between \( y \) and \( u \) by \( \|y - x\| + \|x - s\| + \|s - u\| \), where, using 3,

\[
\|x - s\| = \|x - Q\| - \|s - Q\| = \tau \sin(\alpha/2) - (\tau - \sigma) \sin(\alpha'/2)
\]

and using 4,

\[
\|s - u\| \leq \frac{\delta}{2 \sin(\alpha'/4)}.
\]
Figure 4: Left: edges $e_1$ and $e_2$ with minimum condition number $\tau$ forming the smallest angles $\alpha$ at vertex $x$. Right: The distance $\|F - G\|$ between the two connected components of the shell around an edge point $y'$ must be greater than $\delta$.

Therefore we require that $r$, the inner radius of the shell at Step 4 satisfies

$$r \geq \frac{\delta}{2} + \|x - s\| + \frac{\delta}{2 \sin(\alpha'/4)} =: r_1. \quad (9)$$

Another condition on $r$ arises when we label edge points far from actual vertices. See Figure 4 (right). The distance $\|F - G\|$ between the two connected components of the shell around an edge point $y'$ must be greater than $\delta$. By the cosine law, $\cos(\beta) = \frac{(\tau + \sigma)^2 + (\tau - \sigma)^2 - r^2}{2(\tau^2 - \sigma^2)}$ and

$$\|F - G\| = 2(\tau - \sigma) \sin(\beta) = 2(\tau - \sigma) \sin \left( \arccos \frac{2(\tau^2 + \sigma^2) - r^2}{2(\tau^2 - \sigma^2)} \right) = 2(\tau - \sigma) \sqrt{1 - \left( \frac{2(\tau^2 + \sigma^2) - r^2}{2(\tau^2 - \sigma^2)} \right)^2}.$$

Therefore the condition $\|F - G\| \geq \delta$ can be written as

$$r \geq \sqrt{2(\tau^2 + \sigma^2) - 2(\tau^2 - \sigma^2) \sqrt{1 - \left( \frac{\delta}{2(\tau - \sigma)} \right)^2}} =: r_2. \quad (10)$$

By (9) and (10) we set $r = \max(r_1, r_2)$.

The outer radius of the shell at Steps 3-4 has length $r + \delta$. This guarantees that when the shell around an edge point intersects the tube around $G$ there is at least a point $y \in \mathcal{Y}$ in each connected component of the shell, since $\mathcal{Y}$ is $\frac{\delta}{2}$-dense in $G \oplus \sigma$. 

11
Then, for any edge of $G$ we want to be sure that there is at least one point in $\mathbb{Y}$ which is labeled as an edge point. We start by considering two cases, depicted in Figure 5:

(a) The segment of length $r + \delta$ orthogonal to the face of the tube around edge $e_1$ intersects the face of the tube around edge $e_2$ at a point $z$. This happens when $\alpha' \leq \frac{\pi}{2}$ or when $\delta + r \geq (\tau - \sigma) \frac{\sin(2\alpha' - \pi)}{\sin(\alpha')}$.

(b) There are no segments of length $r + \delta$ orthogonal to the face of the tube around edge $e_1$ that intersect the tube around edge $e_2$. This happens when $\alpha' > \frac{\pi}{2}$ and $\delta + r \leq (\tau - \sigma) \frac{\sin(2\alpha' - \pi)}{\sin(\alpha')}$. In this case we simply consider the segment of length $r + \delta$ from $s$ to a point $z$ on $e_2$.

Suppose $z \in \mathbb{Y}$. Among the points that might be labeled as vertices at Step 6 because of their closeness to vertex $x$, $z$ is the furthest from $x$, since the shell around $z$ is tangent to the tube around $e_1$. At Step 11, in order to control the labelling of the points in the tube between $y$ and $z$ we would like to label all the points in $\{ y' \in \mathbb{Y} : \| y' - y \| \leq \| y - z \| \}$ as vertices. To simplify the calculation we use the following bound

$$\| y - z \| \leq \| y - x \| + \| x - s \| + \| s - z \|$$

where, using (11),

$$\| s - z \| \leq \tilde{s}z := \begin{cases} \frac{r + \delta}{\sin(\alpha'/4)} & \text{case (a)} \\ r + \delta & \text{case (b)} \end{cases}$$

Therefore we set $p_{11} := \frac{\delta}{2} + \| x - s \| + \tilde{s}z$ and at Step 11 we label all the points in $\{ y' \in \mathbb{Y} : \| y' - y \| \leq p_{11} \}$ and $y$ is labeled as vertex at Step 6 \} as vertices.

If $z$ is actually labeled as a vertex at Step 6, then through the expansion of Step 11, all the points at distance not greater than $p_{11}$ from $z$ are labeled as vertices, too.

Finally we determine the length of edge $e_2$ so that there is at least a point in the tube around $e_2$ labeled as an edge point after Step 11. Consider the worst case in which $e_1$ and $e_2$ are forming an angle of size $\alpha$ at both their extremes $x$ and $x'$. See Figure 6.
Figure 6: Edges $e_1$ and $e_2$, forming an angle of size $\alpha$ at both their extremes $x$ and $x'$.

All the points $y' \in \mathcal{Y}$ such that $\|y' - z\| \leq p_{11}$ or $\|y' - z'\| \leq p_{11}$ might be labeled as vertices. When we construct $\mathcal{R}(\mathcal{E})_\delta$ and $\mathcal{R}(\mathcal{V})_\delta$ at Step 15 the two sets of vertices on $e_2$ must be disconnected and there must be at least an edge point on $e_2$. A sufficient condition is that the length of edge $e_2$ is greater than $2(a_1 + a_2 + a_3) + a_4$, where

- $a_1$ is the length of the arc of $e_2$ formed by the projections of lines $\bar{O}x$ and $\bar{O}s$ on $e_2$,
- $a_2$ is the length of the arc of $e_2$ formed by the projection of the chord of length $\hat{s}z$,
- $a_3$ is the length of the arc of $e_2$ formed by the projection of the chord of length $p_{11}$,
- $a_4$ is the length of the arc of $e_2$ formed by the projection of the chord of length $\delta$.

Note that $e_2 = 2\tau \arcsin\left(\frac{\|x-x'\|}{2\tau}\right) = \alpha \tau$ but in general the shortest edge $b$ might be shorter than it. So we require

$$\min(b, \alpha \tau) > 2(a_1 + a_2 + a_3) + a_4. \quad (12)$$

By simple properties involving arcs and chords we have

$$a_1 = \left(\frac{\alpha - \alpha'}{2}\right) \tau, \quad a_2 = 2\tau \arcsin\left(\frac{\hat{s}z}{2(\tau - \sigma)}\right), \quad a_3 = 2\tau \arcsin\left(\frac{p_{11}}{2(\tau - \sigma)}\right), \quad a_4 = 2\tau \arcsin\left(\frac{\delta}{2(\tau - \sigma)}\right).$$
Since the arcsin is superadditive in [0, 1] we require the stronger condition
\[ \min(b, \alpha \tau) - (\alpha - \alpha') \tau > 2\tau \arcsin \left( \frac{2\tilde{z} + 2p_{11} + \delta}{2(\tau - \sigma)} \right), \]
which holds if
\[ \sin \left( \frac{\min(b, \alpha \tau) - (\alpha - \alpha') \tau}{2\tau} \right) > \frac{2\tilde{z} + 2p_{11} + \delta}{2(\tau - \sigma)}. \]
If this condition is satisfied then the graph is correctly reconstructed at Steps 15-17: every connected component of \( R_\delta(V) \) corresponds to a vertex of \( G \) and every connected component of \( R_\delta(E) \) corresponds to an edge of \( G \).

**Analysis of Algorithm 1 with Rips graph distances.**

So far we have applied the reconstruction algorithm using the Euclidean distance at every step. Now we explore the consequences of using a different metric, a strategy closer to the original idea of Aanjaneya et al. (2012). In general, we obtain a less strict condition on \( \delta \) expressed through a more complicated expression than the one of Proposition 5. Let \( \kappa > 0 \) and consider the Rips-Vietoris graph \( R_\kappa(Y) \) with vertex set \( Y \) and edges connecting all the pairs of vertices at distance not greater than \( \kappa \) from each other in \( R_2 \). The metric \( d_Y \) is defined as the shortest path distance on \( R_\kappa(Y) \). Note that the distance \( d_Y \) is used at Steps 3-5, while at Step 11 we still consider the Euclidean distance for the expansion. The following result is just a version of Theorem 2 in Bernstein et al. (2000).

**Lemma 6** Let \( \kappa \) and \( \delta \) be positive, with \( 2\delta \leq \kappa \). If \( Y = \{y_1, \ldots, y_n\} \) is \( \delta \)-dense in \( G \oplus \sigma \), then for all pairs of data points \( y, y' \in G \oplus \sigma \) we have
\[ d_Y(y, y') \leq \left( 1 + \frac{2\delta}{\kappa} \right) d_{G\oplus\sigma}(y, y'), \]
where \( d_{G\oplus\sigma}(y, y') \) is the length of the shortest piecewise-smooth path connecting \( y \) and \( y' \) in the Euclidean space, such that it is entirely contained in \( G \oplus \sigma \).

The parameters \( r \) and \( p_{11} \) are now defined as follows:
\[
 r = \max \left\{ \left( 1 + \frac{2\delta}{\kappa} \right) \left( \frac{\delta}{2} + \tau \sin(\alpha/2) - (\tau - \sigma) \sin(\alpha'/2) + 2\tau \arcsin \left( \frac{\delta}{4(\tau - \sigma) \sin(\alpha'/4)} \right) \right), \right. \\
\left. \left( 1 + \frac{2\delta}{\kappa} \right \sqrt{2(\tau^2 + \sigma^2) - 2(\tau^2 - \sigma^2)} \sqrt{1 - \left( \frac{\delta}{2(\tau - \sigma)} \right)^2} \right\}, \right.
\]
and
\[
 p_{11} = \begin{cases} 
 \frac{3\delta}{2} + \tau \sin(\alpha/2) - (\tau - \sigma) \sin(\alpha'/2) + r & \text{if } \alpha' \geq \pi/2 \text{ and } \kappa \leq (\tau - \sigma) \frac{\sin(2\alpha' - \pi)}{\sin(\alpha')} \\
 \frac{3\delta}{2} + \tau \sin(\alpha/2) - (\tau - \sigma) \sin(\alpha'/2) + \frac{\kappa}{\sin(\alpha'/4)} + r - \kappa & \text{otherwise.} 
\end{cases}
\]
This choice is justified in the proof of Proposition 7.
Proposition 7 If \( \mathcal{Y} \) is \( \frac{\delta}{2} \)-dense in \( G \oplus \sigma \) and

\[
0 \leq \sigma < \tau (1 - \cos(\alpha/2)),
\]

\[
2\delta \leq \kappa < \xi - 2\sigma,
\]

\[
\sin \left( \frac{\min(b, \alpha \tau) - (\alpha - \alpha') \tau}{2\tau} \right) > \begin{cases} \frac{p_1 + 3\delta/2 + r}{\tau - \sigma} & \text{if } \alpha' \geq \frac{\pi}{2} \text{ and } \\ \frac{\kappa}{\sin(\alpha'/4)} + \frac{p_1 + 3\delta/2 + r - \kappa}{\tau - \sigma} & \text{otherwise} \end{cases}
\]

then the graph provided by Algorithm 1 (input: \( \mathcal{Y}, d_\mathcal{Y} \)) is isomorphic to \( G \).

**Proof** The steps are similar to those in the proof of Proposition 5. The condition \( \sigma < \tau (1 - \cos(\alpha/2)) \) is unchanged. The condition \( \kappa < \xi - 2\sigma \) has the same meaning of condition \( (6) \). Now we define the inner radius of the shell. See Figure 4. Since at Steps 3-5 we use the distance \( d_\mathcal{Y} \) on the Rips graph \( R_\kappa \), we need an upper bound on shortest path connecting \( y \) and \( u \) inside the tube. This is given by \( \|y - x\| + \|x - s\| + \tilde{s}u \), where the last term is the length of the arc connecting \( s \) and \( u \); \( \tilde{s}u = 2(\tau - \sigma) \arcsin \frac{\|s - u\|}{2(\tau - \sigma)} \). Therefore, using \( (8) \) and Lemma 6, condition \( (9) \) becomes

\[
r \geq \left( 1 + \frac{2\delta}{\kappa} \right) \left( \frac{\delta}{2} + \|x - s\| + 2(\tau - \sigma) \arcsin \left( \frac{\delta}{4(\tau - \sigma) \sin(\alpha'/4)} \right) \right) =: r_1.
\]

Similarly condition \( (10) \) becomes

\[
r \geq \left( 1 + \frac{2\delta}{\kappa} \right) \sqrt{2(\tau^2 + \sigma^2) - 2(\tau^2 - \sigma^2) \sqrt{1 - \left( \frac{\delta}{2(\tau - \sigma)} \right)^2}} =: r_2
\]

and we set \( r = \max\{r_1, r_2\} \).

Then consider the following two cases, depicted in Figure 7:

(a) The segment of length \( \kappa \) orthogonal to the face of the tube around edge \( e_1 \) intersects the face of the tube around edge \( e_2 \) at a point \( w \). This happens when \( \alpha' \leq \frac{\pi}{2} \) or when \( \kappa \geq (\tau - \sigma) \frac{\sin(2\alpha' - \pi)}{\sin(\alpha')} \).

(b) There are no segments of length \( \kappa \) orthogonal to the face of the tube around edge \( e_1 \) that intersect the tube around edge \( e_2 \). This happens when \( \alpha' > \frac{\pi}{2} \) and \( \kappa \leq (\tau - \sigma) \frac{\sin(2\alpha' - \pi)}{\sin(\alpha')} \). In this case we simply consider the segment of length \( \kappa \) from \( s \) to a point \( w \) on \( e_2 \).

Then we consider point \( z \) on the face of the tube around edge \( e_2 \) such that \( \|w - z\| = r + \delta - \kappa \). Suppose \( z \in \mathcal{Y} \). Among the points that might be labeled as vertices at Step 6 because of their closeness to vertex \( x \), \( z \) is the furthest from \( x \), since the shell \( (d_\mathcal{Y} \text{ distance}) \) around \( z \) is tangent to the tube around \( e_1 \). At Step 11, in order to control the points in the
tube between \(y\) and \(z\) we label all the points in \(\{y' \in Y : \|y' - y\| \leq \|y - z\|\}\) as vertices. To simplify the calculation we use the following bound

\[
\|y - z\| \leq \|y - x\| + \|x - s\| + \|s - w\| + \|w - z\|
\]

\[
\leq \frac{\delta}{2} + \|x - s\| + \|s - w\| + r + \delta - \kappa
\]

where, similarly to (11)

\[
\|s - w\| \leq \tilde{sw} := \begin{cases} \frac{\kappa}{\sin(\alpha'/4)} & \text{case (a)} \\ \kappa & \text{case (b)} \end{cases}
\]

Therefore we set \(p_{11} := \frac{\delta}{2} + \|x - s\| + \tilde{sw} + r + \delta - \kappa\) and at Step 11 we label all the points in \(\{y' \in Y : \|y' - y\| \leq p_{11}\}\) and \(y\) is labeled as vertex at Step 6 as vertices.

Finally we determine the length of edge \(e_2\) so that there is at least a point in the tube around \(e_2\) labeled as an edge point after Step 11. Consider the case of Figure 8 in which \(e_2\) is forming an angle of size \(\alpha\) at both its extremes \(x\) and \(x'\).

All the points \(y' \in Y\) such that \(\|y' - z\| \leq p_{11}\) or \(\|y' - z'\| \leq p_{11}\) might be labeled as vertices. When we construct \(R(E)_\delta\) and \(R(V)_\delta\) at Step 15 the two sets of vertices on \(e_2\) must be disconnected and there must be at least an edge point on \(e_2\). A sufficient condition is that the length of edge \(e_2\) is greater than \(2(a_1 + a_2 + a_3 + a_4) + a_5\), where

- \(a_1\) is the length of the arc of \(e_2\) formed by the projections of lines \(\overline{Ox}\) and \(\overline{Os}\),
- \(a_2\) is the length of the arc of \(e_2\) formed by the projection of the chord of length \(\tilde{sw}\),
- \(a_3\) is the length of the arc of \(e_2\) formed by the projection of the chord of length \(r + \delta - \kappa\),
- \(a_4\) is the length of the arc of \(e_2\) formed by the projection of the chord of length \(p_{11}\),
- \(a_5\) is the length of the arc of \(e_2\) formed by the projection of the chord of length \(\delta\).
Figure 8: Edges $e_1$ and $e_2$, forming an angle of size $\alpha$ at both their extremes $x$ and $x'$. 

Note that $e_2 = 2\tau \arcsin \left( \frac{\|x-x'\|}{2\tau} \right) = \alpha \tau$ but in general the shortest edge $b$ might be shorter than it. So we require

\[
\min(b, \alpha \tau) > 2(a_1 + a_2 + a_3 + a_4 + a_5). \tag{18}
\]

By simple properties involving arcs and chords we have

\[
a_1 = \left( \frac{\alpha - \alpha'}{2} \right) \tau, \quad a_2 = 2\tau \arcsin \left( \frac{s\hat{w}}{2(\tau - \sigma)} \right), \quad a_3 = 2\tau \arcsin \left( \frac{r + \delta - \kappa}{2(\tau - \sigma)} \right),
\]

\[
a_4 = 2\tau \arcsin \left( \frac{p_{11}}{2(\tau - \sigma)} \right), \quad a_5 = 2\tau \arcsin \left( \frac{\delta}{2(\tau - \sigma)} \right).
\]

Since the $\arcsin$ is superadditive in $[0, 1]$ we require the stronger condition

\[
\min(b, \alpha \tau) - (\alpha - \alpha')\tau > 2\tau \arcsin \left( \frac{2s\hat{w} + 2(r + \delta - \kappa) + 2p_{11} + \delta}{2(\tau - \sigma)} \right)
\]

which holds if

\[
\sin \left( \frac{\min(b, \alpha \tau) - (\alpha - \alpha')\tau}{2\tau} \right) > \frac{2s\hat{w} + 2p_{11} + 3\delta + 2(r - \kappa)}{2(\tau - \sigma)}.
\]
If this condition is satisfied then the graph is correctly reconstructed at Steps 15-17: every connected component of $R_\delta(V)$ corresponds to a vertex of $G$ and every connected component of $R_\delta(E)$ corresponds to an edge of $G$.

**Remark 8** We believe that Propositions 5 and 7 are still valid in higher dimensions. That is, the worst case treated in $\mathbb{R}^2$ is the worst case even in higher dimensions.

4. Minimax Analysis

In this section we derive lower and upper bound for the minimax risk

$$R_n = \inf_{\hat{G}} \sup_{P \in \mathcal{P}} P^n \left( \hat{G} \not\simeq G \right),$$

where, as described in Section 2, the infimum is over all estimators $\hat{G}$ of the metric graph $G$, the supremum is over the class of distributions $\mathcal{P}$ for $\mathcal{Y}$ and $\hat{G} \not\simeq G$ means that $\hat{G}$ and $G$ are not isomorphic.

4.1 Lower Bound

To derive a lower bound on the minimax risk, we make repeated use of Le Cam’s lemma. See, e.g., Tsybakov (2008). Recall that the total variation distance between two measures $P$ and $Q$ on the same probability space is defined by $\text{TV}(P, Q) = \sup_A |P(A) - Q(A)|$ where the supremum is over all measurable sets.

**Lemma 9 (Le Cam)** Let $\mathcal{Q}$ be a set of distributions. Let $\theta(Q)$ take values in a metric space with metric $\rho$. Let $Q_1, Q_2 \in \mathcal{Q}$ be any pair of distributions in $\mathcal{Q}$. Let $Y_1, \ldots, Y_n$ be drawn iid from some $Q \in \mathcal{Q}$ and denote the corresponding product measure by $Q^n$. Then

$$\inf_{\hat{\theta}} \sup_{Q \in \mathcal{Q}} \mathbb{E}_{Q^n} \left[ \rho(\hat{\theta}, \theta(Q)) \right] \geq \frac{1}{8} \rho(\theta(Q_1), \theta(Q_2))(1 - \text{TV}(Q_1, Q_2))^{2n}$$

where the infimum is over all the estimators.

Below we will apply Le Cam’s lemma using several pairs of distributions. Any pair $Q_1, Q_2$ will be associated with a pair of metric graphs $G_1, G_2 \in \mathcal{G}$. We will take $\theta(Q_1)$ and $\theta(Q_2)$ to be the classes of graphs that are isomorphic to $G_1$ and $G_2$. We will set $\rho(\theta(Q_1), \theta(Q_2)) = 0$ if $G_1$ and $G_2$ are isomorphic and $\rho(\theta(Q_1), \theta(Q_2)) = 1$ otherwise.

**Theorem 10** In the noiseless case, for $b \leq b_0(a)$, $\alpha \leq \alpha_0(a)$, $\xi \leq \xi_0(a)$, $\tau \leq \tau_0(a)$, where $b_0(a), \alpha_0(a), \xi_0(a)$ and $\tau_0(a)$ are constants which depend on $a$, a lower bound on the minimax risk for metric graph reconstruction is

$$R_n \geq \frac{1}{8} \exp \left( -2a \min \{ b, 2\sin(\alpha/2), \xi, 2\pi\tau \} n \right).$$
Proof We consider the 4 parameters separately.

Shortest edge $b$
Consider the metric graph $G_1$ consisting of a single edge of length $1+b$ and another metric graph $G_2$ with an edge of length 1 and an orthogonal edge of length $b$ glued in the middle. See Figure 9. The density on $G_1$ is constructed in the following way: on the set $W_1 = G_1 \setminus G_2$ of length $b$ we set $p_1(x) = a$ and the rest of the mass is evenly distributed over the remaining portion of $G_1$. Similarly, for $G_2$ we set $p_2(x) = a$ on $W_2 = G_2 \setminus G_1$, which correspond to the orthogonal edge of length $b$. We evenly spread the remaining mass. The two densities differ only on the sets $W_1$ and $W_2$. Therefore

$$TV(p_1, p_2) \leq ab$$

and, by Le Cam’s lemma,

$$R_n \geq \frac{1}{8} (1 - ab)^{2n} \geq \frac{1}{8} e^{-2abn}$$

for all $b \leq b_0(a)$, where $b_0(a)$ is a constant depending on $a$.

Smallest angle $\alpha$
Now consider the metric graphs of Figure 10. $G_3$ consists of two edges of length 2 forming an angle $\alpha$ and a third edge of length $1 + 2 \sin(\alpha/2)$ glued to the first two. $G_4$ is similar: an edge of length $2 \sin(\alpha/2)$ is added to complete the triangle, while the edge on the left has length 1.

As in the previous case we set $p_3(x) = a$ on $W_3 = G_3 \setminus G_4$, $p_4(x) = a$ on $W_4 = G_4 \setminus G_3$ and spread evenly the rest of the mass. The total variation distance is

$$TV(p_3, p_4) \leq 2a \sin\left(\frac{\alpha}{2}\right)$$
and, by Le Cam’s lemma,

\[ R_n \geq \frac{1}{8}(1 - 2a \sin(\alpha/2'))^{2n} \geq \frac{1}{8}e^{-4a\sin(\alpha/2)n} \]

for all \( \alpha \leq \alpha_0(a) \), where \( \alpha_0(a) \) is a constant depending on \( a \).

**Global condition number \( \xi \)**

We defined the global condition number as the shortest euclidean distance between two points far away in the graph distance. In Figure 11, we have metric graph \( G_5 \) formed by a single edge of length 1 and metric graph \( G_6 \) consisting of two edges of length 0.5, \( \xi \) apart from each other.

Figure 11: \( G_5 \) and \( G_6 \)

Again, we set \( p_5(x) = a \) on \( W_5 = G_5 \setminus G_6 \), \( p_6(x) = a \) on \( W_6 = G_6 \setminus G_5 \); and evenly spread the rest. We obtain

\[ TV(p_5, p_6) \leq a\xi \]

and, by Le Cam’s lemma,

\[ R_n \geq \frac{1}{8}(1 - a\xi')^{2n} \geq \frac{1}{8}e^{-2a\xi n} \]

for all \( \xi \leq \xi_0(a) \), where \( \xi_0(a) \) is a constant depending on \( a \).

**Local condition number \( \tau \)**

The local condition number \( \tau \) is the smallest condition number of the edges forming the metric graph. In Figure 12, we have metric graph \( G_7 \) consisting of a loop of radius \( \tau \) attached to an edge of length 1 and metric graph \( G_8 \), a single edge of length \( 1 + \pi\tau \).

Figure 12: \( G_7 \) and \( G_8 \)

As in the previous cases \( p_7(x) = a \) on \( W_7 = G_7 \setminus G_8 \) and \( p_8(x) = a \) on \( W_8 = G_8 \setminus G_7 \). It follows that

\[ TV(p_7, p_8) \leq 2a\pi\tau \]
and, by Le Cam’s lemma,

\[ R_n \geq \frac{1}{8}(1 - 2a\pi\tau)^{2n} \geq \frac{1}{8}e^{-4a\pi\tau n} \]

for all \( \tau \leq \tau_0(a) \), where \( \tau_0(a) \) is a constant depending on \( a \).

### 4.2 Upper Bound

In this section we use the analysis of the performance of Algorithm 1 to derive an upper bound on the minimax risk. We first need two useful lemmas.

**Lemma 11 (5.1 in Niyogi et al. (2008))** Let \( \{A_i\} \) for \( i = 1, \ldots, l \) be a finite collection of measurable sets and let \( \mu \) be a probability measure on \( \bigcup_{i=1}^{l} A_i \) such that for all \( 1 \leq i \leq l \), we have \( \mu(A_i) > \gamma \). Let \( \bar{x} = \{x_1, \ldots, x_n\} \) be a set of \( n \) i.i.d. draws according to \( \mu \). Then if

\[ n \geq \frac{1}{\gamma} \left( \log l + \log \left( \frac{1}{\lambda} \right) \right) \]

we are guaranteed that with probability \( > 1 - \lambda \), the following is true:

\[ \forall i, \; \bar{x} \cap A_i \neq \emptyset. \]

Recall that the *covering number* \( C(\delta) \) of a set is the smallest number of balls of radius \( \delta \) required to cover the set. The *packing number* \( P(\delta) \) is the largest number of non-overlapping balls of radius \( \delta \) whose centers are contained in the set.

**Lemma 12 (5.2 in Niyogi et al. (2008))** For every \( \delta > 0 \),

\[ P(2\delta) \leq C(2\delta) \leq P(\delta). \]

By restricting the attention to metric graphs embedded in \( \mathbb{R}^2 \) and using Proposition 5, we obtain the following upper bound on \( R_n \).

**Theorem 13** In the noiseless case \( (\sigma = 0) \), for \( \alpha \leq \pi/2 \), an upper bound on the minimax risk \( R_n \) is given by

\[ R_n \leq \frac{8 \text{length}(G)}{\delta} \exp \left\{ -\frac{a\delta}{4 \text{length}(G)} \right\}, \]

where

\[ \delta = \frac{1}{2} \min \left\{ \xi, \frac{2\sin(\alpha/4)}{3\sin(\alpha/4) + 1}, \frac{\tau\sin^2(\alpha/4)}{\sin^2(\alpha/4) + 3\sin(\alpha/4) + 1} \sin \left( \frac{\min\{b, \alpha\tau\}}{2\tau} \right) \right\}. \]
Proof In the noiseless case, for $\alpha \leq \pi/2$, Proposition 5 implies that the graph $G$ can be reconstructed from a $\delta/2$–dense sample $Y$ if

$$\delta < \min \left\{ \frac{2\sin(\alpha/4)}{3\sin(\alpha/4) + 1}, \frac{\tau\sin^2(\alpha/4)}{\sin^2(\alpha/4) + 3\sin(\alpha/4) + 1} \sin \left( \frac{\min\{b, \alpha \tau\}}{2\tau} \right) \right\}. \quad (22)$$

Condition (22) follows from (6)-(7) and the simplified forms of $r$ and $p_{11}$ for the noiseless case. We simply choose $\delta = \frac{1}{2} \min \left\{ \frac{2\sin(\alpha/4)}{3\sin(\alpha/4) + 1}, \frac{\tau\sin^2(\alpha/4)}{\sin^2(\alpha/4) + 3\sin(\alpha/4) + 1} \sin \left( \frac{\min\{b, \alpha \tau\}}{2\tau} \right) \right\}$ and we look for the sample size $n$ that guarantees a $\delta/2$–dense sample with high probability.

Following the strategy in Niyogi et al. (2008), we consider a cover of the metric graph $G$ by balls of radius $\delta/4$. Let $\{x_i : 1 \leq i \leq l\}$ be the centers of such balls that constitute a minimal cover. We can choose $A_{\delta/4} = B_{\delta/4}(x_i) \cap G$. Applying Lemma 11 we find that the sample size that guarantees a correct reconstruction with probability at least $1 - \lambda$ is

$$\frac{1}{\gamma} \left( \log l + \log \frac{1}{\lambda} \right),$$

where

$$\gamma \geq \min_i \frac{\text{length}(A_{\delta/4})}{\text{length}(G)} \geq \frac{a\delta}{4\text{length}(G)},$$

and we bound the covering number $l$ in terms of the packing number, using Lemma 12

$$l \leq \frac{\text{length}(G)}{\text{length}(A_{\delta/8})} \leq \frac{8\text{length}(G)}{\delta}.$$

Therefore if

$$n = \frac{4\text{length}(G)}{a\delta} \left[ \log \left( \frac{8\text{length}(G)}{\delta} \right) + \log \frac{1}{\lambda} \right] \quad (23)$$

we have $P(\hat{G} \not\cong G) \leq \lambda$. Rearranging we have the desired result.

Note that the upper and lower bounds are tight up to polynomial factors in the parameters $\tau, b, \xi$. There is a small gap with respect to $\alpha$; closing this gap is an open problem.

5. Experiments

Example 2 The Topology of Shadyside. In this example we try to reconstruct the main streets of Shadyside, a neighborhood of Pittsburgh. Figure 13 shows the main streets (yellow) of Shadyside obtained from Google Maps. There are 16 vertices and 21 edges for a total length of 9,725 m. We imagine taking a walk in Shadyside and recording our position at several points in order to reconstruct the topology of the graph formed by the streets. Propositions 5 and 7 tell us how often we should record our position. The shortest edge has length 120 m and the smallest angle is $\pi/3$ (at Highland & Centre Ave). The streets...
are mostly straight, with a few exceptions. We assume $\tau = 300m$. Also, $\xi = 90m$ since Ellsworth Ave and Centre Ave become very close in a point far from the vertices.

Assuming no noise, conditions (6)–(7) simplify to (22) from which we get $\delta \leq 2.17$. On the other hand, condition (15) is satisfied for $\delta = 2.46$ and $\kappa = 2\delta$. Therefore registering the position every 2.46 m would guarantee a correct reconstruction. Algorithm 1 provides the estimated graph plotted in red on top of the map in Figure 13. By (23), if we sample points uniformly we need 229,273 points to reconstruct the graph with probability at least 0.9.

### Example 3  A Neuron in Three-Dimensions.

In support of Remark 8, we return to the neuron example and we try to apply Propositions 5 and 7 to the 3D data of Figure 1, namely the neuron cr22e from the hippocampus of a rat (Gulyás et al., 1999). The data were obtained from NeuroMorpho.Org (Ascoli et al., 2007). The total length of the graph is $1750.86\mu m$. We assume the smallest edge has length $100\mu m$, the smallest angle $\pi/3$, the local condition number 30 and $\xi = 50$. The conditions of Proposition 5 are satisfied for $\delta = 0.54$, while those of Proposition 7 are satisfied for $\delta = 0.62$ and $\kappa = 2\delta$. Figure 1b shows the reconstructed graph. If we sample points uniformly from the original metric graph we need 114,435 points to correctly reconstruct the graph with probability at least 0.9.

### 6. Conclusion

In this paper, we presented a statistical analysis of metric graph reconstruction. We derived sufficient conditions on random samples from a graph metric space that guarantee topological reconstruction and we derived lower and upper bounds on the minimax risk for this problem.

Various improvements and theoretical extensions are possible. First, we conjecture that our analysis for the performance of the algorithm of Aanjaneya et al. (2012) under the choice of parameters we describe and the corresponding minimax rates will extend to
higher dimensions. Currently, we are investigating the idea of combining metric graph reconstruction with the subspace constrained mean-shift algorithm (Fukunaga and Hostetler 1975; Comaniciu and Meer 2002; Genovese et al. 2012) to provide similar guarantees. Our preliminary results indicate that this mixed strategy works very well under more general noise assumptions and with relatively low sample size.

Acknowledgments

We would like to acknowledge support for this project from NSF CAREER grant DMS 114967.

References

Mridul Aanjaneya, Frederic Chazal, Daniel Chen, Marc Glisse, Leonidas Guibas, and Dmitriy Morozov. Metric graph reconstruction from noisy data. International Journal of Computational Geometry &amp; Applications, 22(04):305–325, 2012.

Mahmuda Ahmed and Carola Wenk. Probabilistic street-intersection reconstruction from gps trajectories: approaches and challenges. In Proceedings of the Third ACM SIGSPATIAL International Workshop on Querying and Mining Uncertain Spatio-Temporal Data, pages 34–37. ACM, 2012.

Ery Arias-Castro, Guangliang Chen, and Gilad Lerman. Spectral clustering based on local linear approximations. Electronic Journal of Statistics, 5:1537–1587, 2011.

Giorgio A Ascoli, Duncan E Donohue, and Maryam Halavi. Neuromorpho.org: a central resource for neuronal morphologies. The Journal of Neuroscience, 27(35):9247–9251, 2007.

Paul Bendich. Analyzing stratified spaces using persistent versions of intersection and local homology. ProQuest, 2008.

Paul Bendich, Sayan Mukherjee, and Bei Wang. Towards stratification learning through homology inference. arXiv preprint arXiv:1008.3572, 2010.

Paul Bendich, Bei Wang, and Sayan Mukherjee. Local homology transfer and stratification learning. In Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms, pages 1355–1370. SIAM, 2012.

Mira Bernstein, Vin De Silva, John C Langford, and Joshua B Tenenbaum. Graph approximations to geodesics on embedded manifolds. Technical report, Technical report, Department of Psychology, Stanford University, 2000.

F. Chazal and A. Lieutier. Topology guaranteeing manifold reconstruction using distance function to noisy data. In Proceedings of the twenty-second annual symposium on Computational geometry, pages 112–118. ACM, 2006.
Daniel Chen, Leonidas J Guibas, John Hershberger, and Jian Sun. Road network reconstruction for organizing paths. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1309–1320. Society for Industrial and Applied Mathematics, 2010.

Alexey Chernov and Vitaliy Kurlin. Reconstructing persistent graph structures from noisy images. *Electronic Journal Imagen-A*, 3, 2013.

Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(5):603–619, 2002.

Keinosuke Fukunaga and Larry Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *Information Theory, IEEE Transactions on*, 21(1):32–40, 1975.

Christopher R Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Nonparametric ridge estimation. *arXiv preprint arXiv:1212.5156*, 2012.

Attila I Gulyás, Manuel Megías, Zsuzsa Emri, and Tamás F Freund. Total number and ratio of excitatory and inhibitory synapses converging onto single interneurons of different types in the CA1 area of the rat hippocampus. *The Journal of neuroscience*, 19(22):10082–10097, 1999.

Peter Kuchment. Quantum graphs: I. some basic structures. *Waves in Random media*, 14 (1):107–128, 2004.

P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence. *Discrete and Computational Geometry*, 38(1-3):419–441, 2008.

Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer, 2008.