Journal of Machine Learning Research 15 (2014) 3425-3446 Submitted 5/13; Revised 6/14; Published 10/14

Statistical Analysis of Metric Graph Reconstruction

Fabrizio Lecci LECCI@CMU.EDU
Alessandro Rinaldo ARINALDO@CMU.EDU
Larry Wasserman LARRY@CMU.EDU
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Editor: Matthias Hein

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 embedded in $\mathbb{R}^D$ from a random sample. We derive lower and upper bounds on the minimax risk for the noiseless case and tubular noise case. 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. Data sets 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 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. Finally, we obtain a lower bound on the risk of topological reconstruction, which, in the noiseless case, 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. In Section 5 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. From the theoretical side, we replace the key assumption in Aanjaneya et al. (2012) of the sample being a (ε, R)-approximation to the underlying metric graph, by the milder assumption of the sample being dense in a neighborhood of the metric graph. Approximation and reconstruction of metric graphs has also been considered in Chazal and Sun (2013) and Ge et al. (2011). 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). 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. The authors assume that the sample comes with a metric that is close to the intrinsic metric of the underlying graph, by imposing a limit on the Gromov-Hausdorff distance between the two metrics. By considering data embedded in the Euclidean space and focusing on the topological aspect, we show that the notion of *dense sample* is sufficient to guarantee a correct reconstruction.

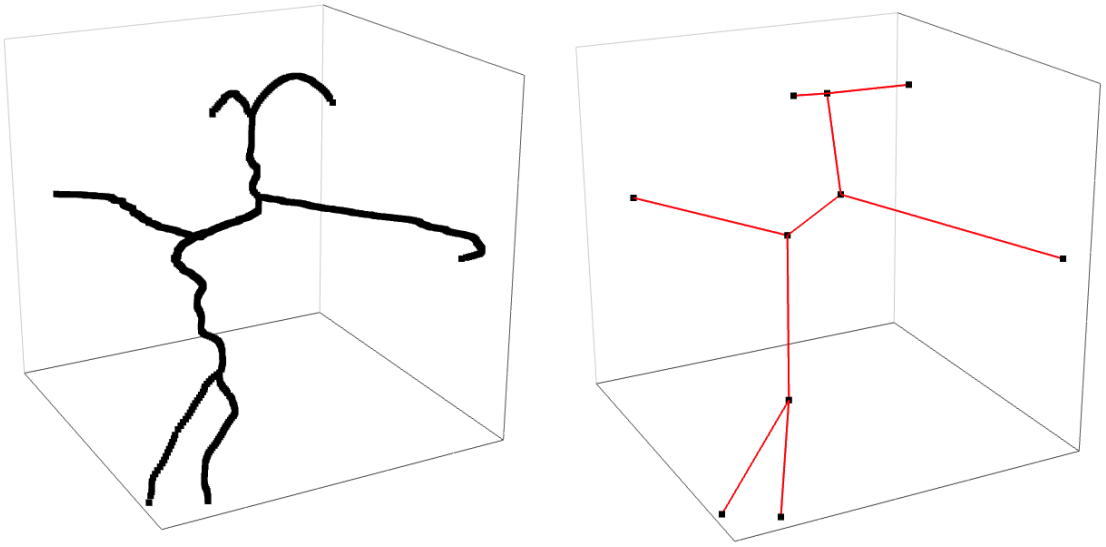

Figure 1: Left: Neuron cr22e from the hippocampus of a rat; NeuroMorpho.Org (Ascoli et al., 2007). Right: A metric graph reconstruction of the neuron.

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 $u = v$).

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 definition of reach. Let M be a 1-dimensional manifold embedded in $\mathbb{R}^D$. Let $T_u M$ denote the 1-dimensional tangent space to M and let $T_u^{\perp}\mathrm{M}$ be the $(D-1)$-dimensional normal space.

Definition 2 *Define the fiber of size a at $u \in M$ to be $L_a(u, M) = T_u^{\perp} M \bigcap B(u, a)$, where $B(u, a)$ is the D-dimensional ball of radius a centered at u. If M has boundary $\{v_1, v_2\}$, the fiber of size a at v_i is defined as the limit of $L_a(u, M)$, as u approaches v_i in $M \backslash \{v_1, v_2\}$. The reach of M is the largest number τ such that the fibers $L_\tau(u, M)$ never intersect.*

The reach sets a limit on the curvature of a manifold. A manifold with large reach does not come too close to be self-intersecting. For example the reach of an arc of a circle is equal to its radius. The quantity $1/\tau$ is called the *condition number* in Niyogi et al. (2008). For more details see also Federer (1959); Chazal and Lieutier (2006); Genovese et al. (2012a). Each edge of a metric graph (G, d_G) can be seen as a 1-dimensional manifold with boundary. Let the *local reach* of metric graph G be the minimum reach associated to an edge of G.

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 ϵ centered at x. Let $\ell_1(\epsilon)$ be the line segment joining the two points x and $\partial B(x, \epsilon) \bigcap e_1$. Let $\ell_2(\epsilon)$ be the line segment joining the two points x and $\partial B(x, \epsilon) \bigcap e_2$. 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.

To control points far away in the graph distance, but close in the embedding space, we define

$$A_G = \{(x, x') \in G \times G : d_G(x, x') \geq \min(b, \tau\alpha)\},$$

where b is the shortest edge of G, τ is the local condition number and α is the smallest angle formed by two edges of G. We define the *global reach* as the infimum of the Euclidean distances among pairs of point in A_G, that is $\xi = \inf_{A_G} \|x - x'\|_2$.

Let (G, d_G) be a metric graph and, for a constant $\sigma \geq 0$, let $G_\sigma = \{y : \inf_{x \in G} ||x - y||_2 \leq \sigma\}$ be the σ-tube around G. If $\sigma = 0$, then, trivially, $G_\sigma = G$. Notice that G_σ is a set of dimension D if $\sigma > 0$.

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

Definition 3 *The sample $\mathbb{Y} = \{y_1, \ldots, y_n\} \subset G_\sigma \subset \mathbb{R}^D$ is $\frac{\delta}{2}$-dense in G_σ if for every $x \in G_\sigma$, there exists a $y \in \mathbb{Y}$ such that $\|x - y\|_2 < \frac{\delta}{2}$.*

The problem of metric graph reconstruction consists of reconstructing a metric graph G given a dense sample $\{y_1, \ldots, y_n\} = \mathbb{Y} \subset G_\sigma$ endowed with a distance $d_{\mathbb{Y}}$, which could be the D-dimensional Euclidean distance or some more complicate notion of 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.

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

A1 The graphs have finite total length and 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 reach at least $\tau > 0$.

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

A4 The global reach is at least $\xi > 0$.

Assumptions A1 and A2 allow us to consider each edge of a metric graph as a single smooth curve. A3 and A4 are additional regularity conditions on the separation between different edges. Assumptions similar to A1-A4 are common in the literature. For different regularity conditions that allow for corners within an edge see, for example, Chazal et al. (2009) and Chen et al. (2010).

Let $\mathcal{G}$ be the set of metric graphs embedded in $\mathbb{R}^D$ that satisfy assumptions A1, A2, A3 and A4, involving the parameters b, α, τ, ξ. We consider two noise models:

Noiseless. We observe data $Y_1, \ldots, Y_n \sim P$, where $P \in \mathcal{P}$, a collection 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 $a > 0$.

Tubular Noise. We observe data $Y_1, \ldots, Y_n \sim P_{G,\sigma}$ where $P_{G,\sigma}$ is uniform on the σ-tube G_σ. In this case we consider the collection $\mathcal{P} = \{P_{G,\sigma} : G \in \mathcal{G}\}$.

We are interested in bounding the minimax risk

$$R_n = \inf_{\widehat{G}} \sup_{P \in \mathcal{P}} P^n\Big(\widehat{G} \not\simeq G\Big), \tag{1}$$

where the infimum is over all estimators $\widehat{G}$ of the topology of (G, d_G), the supremum is over the class of distributions $\mathcal{P}$ for $\mathbb{Y}$ and $\widehat{G} \not\simeq G$ means that $\widehat{G}$ and G are not isomorphic.
In Section 4 we will find lower and upper bounds for R_n in the noiseless case and the tubular noise case.
We conclude this section by summarizing the many parameters and symbols involved in our analysis. See Table 1.

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

In this section we study the performance of the metric graph reconstruction algorithm of Aanjaneya et al. (2012), under assumptions A1-A4 and with a choice of parameters adapted to our setting. In Section 4 we will use these results to derive bounds on the minimax rate for topology reconstruction. The metric graph reconstruction algorithm is presented in Algorithm 1.

The algorithm takes a (possibly noisy) sample $\mathbb{Y}$ from a metric graph G and a distance $d_{\mathbb{Y}}$ defined on $\mathbb{Y}$ and returns a graph $\widehat{G}$ that approximates G. The key idea is the following:

Symbol	Meaning
(G, d_G)	metric graph
α	smallest angle
b	shortest edge
τ	local reach
ξ	global reach
$\mathcal{G}$	set of metric graphs embedded in $\mathbb{R}^D$, satisfying A1-A4
$\mathcal{P}$	set of distributions on G or G_σ
G_σ	σ tube around G
$\mathbb{Y}$	sample, subset of G_σ
δ	$\mathbb{Y}$ is a $\delta/2$-dense sample

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

a shell of radius r is constructed around each point in the sample, which is labeled *edge point* if its shell contains 2 well separated clusters of sampled points and *vertex point* otherwise. Several steps of the algorithm require the construction of a Rips-Vietoris graph of parameter δ: $\mathcal{R}_\delta(S_y)$ is a graph whose vertices are all the points of S_y and there is an edge between two points if the Euclidean distance between them is not larger than δ. At Step 11 some of the edge points that are close to vertices are re-labeled as vertex points. This expansion guarantees a precise borderline between clusters of vertex points and clusters of edge points. At steps 15-17 each of these clusters is associated to a vertex or to an edge of the reconstructed graph $\widehat{G}$. We will analyze the algorithm considering the Euclidean

Algorithm 1 Metric Graph Reconstruction Algorithm

Input: sample $\mathbb{Y}, d_{\mathbb{Y}}, r, p_{11}$.
1: **Labeling points as edge or vertex**
2: for all $y \in \mathbb{Y}$ do
3: $S_y \leftarrow B(y, r+\delta) \backslash B(y, r)$
4: $\deg_r(y) \leftarrow$ Number of connected components of Rips-Vietoris graph $\mathcal{R}_\delta(S_y)$
5: if $\deg_r(y) = 2$ 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 $\mathbb{E}$ be the point of $\mathbb{Y}$ labeled as edge points.
13: Let $\mathbb{V}$ be the point of $\mathbb{Y}$ labeled as vertices.
14: **Reconstructing the graph structure**
15: Compute the connected components of the Rips-Vietoris graphs $\mathcal{R}_\delta(\mathbb{E})$ and $\mathcal{R}_\delta(\mathbb{V})$.
16: Let the connected components of $\mathcal{R}_\delta(\mathbb{V})$ be the vertices of of the reconstructed graph $\widehat{G}$.
17. Let there be an edge between vertices of $\widehat{G}$ if their corresponding connected components in $\mathcal{R}_\delta(\mathbb{V})$ contain points at distance less than δ from the same connected component of $\mathcal{R}_\delta(\mathbb{E})$.
Output: $\widehat{G}$.

distance on the sample $\mathbb{Y}$, that is, $d_{\mathbb{Y}} = \|\cdot\|_2$. 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.

Before finding how dense a sample has to be in order to guarantee a correct reconstruction of a metric graph, we show that it is sufficient to study a particular metric graph embedded in $\mathbb{R}^2$, which represents the worst case. In other words, if the metric graph algorithm can reconstruct this particular planar graph, then it can reconstruct any other metric graph that satisfies A1-A4.

3.1 The Worst Case: a Metric Graph in $\mathbb{R}^2$

The worst case is the one for which it is hard to distinguish two edges that intersect at a vertex because they are too close in the embedding space.
Figure 2 (top left) shows an edge e that intersects two edges e_1, e_2 with reach τ, forming an angle α at vertex x. In the plots, the embedding space is $\mathbb{R}^3$ ($D = 3$) and we show the projections of e, e_1 and e_2 on the (limit) plane formed by e_1 and e_2, passing through x.

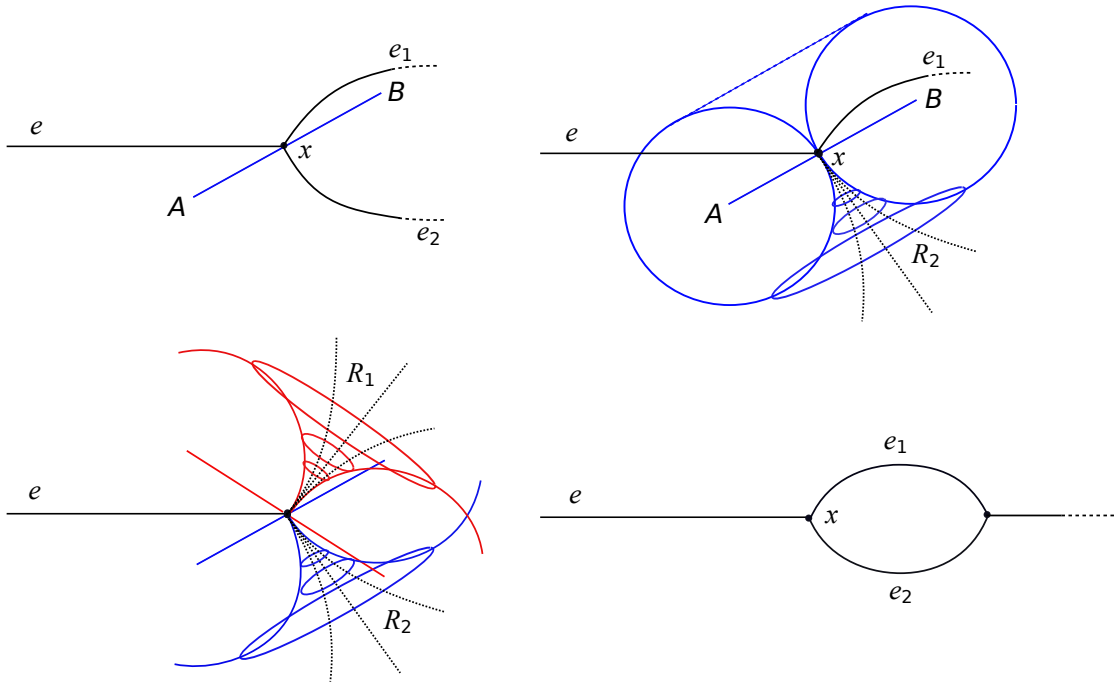

Figure 2: Even in the worst case, edges e_1 and e_2 must lie outside of the torii constructed on the fibers $L_\tau(x, e_1)$ and $L_\tau(x, e_2)$.

We focus on edge e_2. The fiber $L_\tau(x, e_2)$ of size τ around x is a $(D-1)$-dimensional ball centered at x and orthogonal to e_2. In $\mathbb{R}^3$, $L_\tau(x, e_2)$ is a disk of radius τ, whose projection on the plane is the segment $\overline{AB}$. By definition, for any $y \in e_2$, the fiber $L_\tau(y, e_2)$ can not intersect the fiber $L_\tau(x, e_2)$, otherwise the assumption involving the reach τ would be violated. We represent this condition by considering a D-dimensional ball of radius τ, centered at each point of the boundary of $L_\tau(x, e_2)$. Edge e_2 must lie outside of these balls,

in a feasible region that we denote by R_2, so that its fibers do not intersect $L_\tau(x, e_2)$. In $\mathbb{R}^3$, this procedure forms a horn torus (a torus with no hole) around vertex x. See the top right plot of Figure 2. The same reasoning applies to edge e_1, which must lie in the region R_1, outside of the balls of radius τ centered at each point of the boundary of $L_\tau(x, e_1)$. See the bottom left plot.

At each given distance from vertex x, two points of e_1 and e_2 are as close as possible when they lie on the boundaries of R_1 and R_2, on the same (limit) plane formed by e_1 and e_2, passing through x. When e_1 and e_2 lie on this plane, on the boundaries of the two feasible regions, they are as close as possible in the embedding space. This worst case is represented in the bottom right plot of Figure 2. Note that e_1 and e_2 are simply arcs of circles of radius τ.

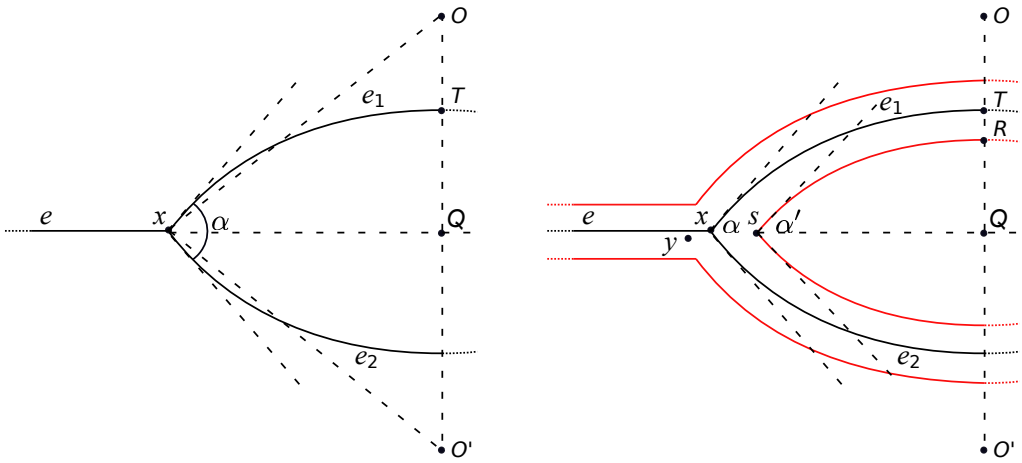

Figure 3: Left: edges e_1 and e_2 with minimum reach τ forming the smallest angles α at vertex x. Right: same metric graph with a tube of radius σ around it.

We will use basic trigonometric properties of the worst case. In Figure 3 (left), 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 $\pi - \alpha$. It can be shown that

$$x\widehat{O}O' = \alpha/2, \tag{2}$$

$$T\hat{x}Q = \alpha/4. \tag{3}$$

Let $\mathbb{Y}$ be a noisy sample of G. In other words $\mathbb{Y}$ is a subset of G_σ, the tube of radius $\sigma \geq 0$ around the metric graph G. See Figure 3 (right). Let Q be the midpoint of segment $\overline{OO'}$ and let T be the intersection point of $\overline{OO'}$ and edge e_1. For $0 \leq \sigma \leq \overline{QT} = \tau - \tau\cos(\alpha/2)$, the smallest angle formed by the inner faces of the tube around the metric graph is

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

where we applied the cosine law to the triangle OsO' and the fact that angle $O\hat{s}O'$ has width $\pi - \alpha'$. Note that if $\sigma = 0$ then $\alpha' = \alpha$. As in (3), it can be shown that

$$R\hat{s}Q = \alpha'/4. \tag{5}$$

The few basic trigonometric equations described above will be used to determine under which conditions on $b, \alpha, \tau, \xi, \sigma$ the metric graph reconstruction algorithm can reconstruct the worst case.

3.2 Analysis of Algorithm 1 with Euclidean Distance

In this section we analyze Algorithm 1. The Euclidean distance is used at every step of the algorithm, which requires the specification of r, the inner radius of the shell, and p_{11}, the parameter governing the expansion of Step 11. We set

$$r = \frac{\delta}{2} + \sigma + \tau \sin(\alpha/2) - (\tau - \sigma)\sin(\alpha'/2) + \frac{\delta}{2\sin(\alpha'/4)} \tag{6}$$

and

$$p_{11} = \frac{\delta}{2} + \tau \sin(\alpha/2) - (\tau - \sigma)\sin(\alpha'/2) + \frac{r + \delta}{\sin(\alpha'/2)} \tag{7}$$

This choice is justified in the proof of Proposition 4.
Define

$$f(b, \alpha, \tau, \xi, \sigma) := \frac{(\tau - \sigma)\sin\left(\frac{\min(b, \alpha\tau) - (\alpha - \alpha')\tau}{2\tau}\right) - [\tau \sin(\alpha/2) - (\tau - \sigma)\sin(\alpha'/2)]\left(1 + \frac{2}{\sin(\alpha'/2)}\right) - \frac{2\sigma}{\sin(\alpha'/2)}}{1 + 3[\sin(\alpha'/2)]^{-1} + [\sin(\alpha'/2)\sin(\alpha'/4)]^{-1}}, \tag{8}$$

where α' is given in 4. Note that $f(b, \alpha, \tau, \xi, \sigma)$ is a decreasing function of σ.

Proposition 4 *If $\mathbb{Y}$ is $\frac{\delta}{2}$-dense in G_σ and*

$$0 < r + \delta < \xi - 2\sigma, \tag{9}$$

$$0 < \delta < f(b, \alpha, \tau, \xi, \sigma), \tag{10}$$

then the graph $\widehat{G}$ provided by Algorithm 1 (input: $\mathbb{Y}, \|\cdot\|_2, r, p_{11}$) is isomorphic to G.

Proof Our objective is to use Algorithm 1 to reconstruct edges and vertices of a metric graph G embedded in R^D. Condition (9) guarantees that points of G which are far apart in the metric graph distance d_G, and close in the embedding space, do not interfere in the construction of the shells at Steps 3-4. Therefore we can restrict the attention to adjacent edges in a neighborhood of the vertex at which they intersect. In particular, since Algorithm 1 is based on the Euclidean distance between the edges, if we can distinguish two adjacent edges that are as close as possible in the embedding space, then we can distinguish any other pair of adjacent edges. As shown in Section 3.1, two adjacent edges e_1 and e_2, forming an angle of width α at vertex x, are as close as possible when they lie on the same plane, on the boundaries of the feasible regions determined by the condition on the reach (assumption A2). We will show that under conditions (9) and (10), Algorithm 1 can reconstruct this worst case. This will imply that the algorithm can reconstruct the topology of other vertices and edges in the D-dimensional space.

The rest of the proof involves condition (10). 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 using Algorithm 1 we want to be sure that y is labeled as a vertex, that is, the number of connected components of the shell around y is different than 2 (Steps 3-4). The worst case is depicted in Figure 4 (left), where x is the vertex of minimum angle α, formed by two edges, e_1 and e_2 of reach τ. First, we show that for the the value of r selected in (6), points close to an actual vertex are labeled as vertices at Steps 3-10 and points far from actual vertices are labeled as edges. The inner faces of the tube of radius σ around e_1 and

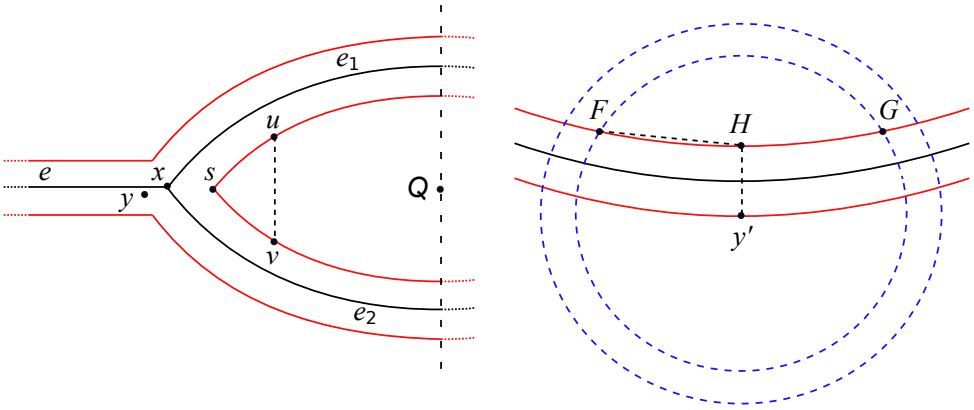

Figure 4: Left: edges e_1 and e_2 with minimum reach τ forming the smallest angles α at vertex x. Right: The distance $\|F - G\|_2$ between the two connected components of the shell around an edge point y' must be greater than δ.

e_2 form an angle of width α' at vertex s, as described in Section 3.1. Let u and v be the two points on the faces of the tube such that they are equidistant from x and $\|u - v\|_2 = \delta$. Since at Step 4 we construct a δ-graph to determine the number of connected components of the shell S_y and we want y to be a vertex, we choose r, the inner radius of the shell S_y, so that if $u, v \in \mathbb{Y}$ then $r \geq \max\{d_{\mathbb{Y}}(y, u), d_{\mathbb{Y}}(y, v)\}$. This guarantees that $\forall t_1, t_2 \in \mathbb{Y}$ with t_1 around edge e_1, t_2 around edge e_2 such that $\{t_1, t_2\} \subset S_y$, we have $d_{\mathbb{Y}}(t_1, t_2) \geq \delta$, that is t_1 and t_2 belong to different connected components of the shell around y at Step 4.
The distance between y and u is bounded by $\|y - x\|_2 + \|x - s\|_2 + \|s - u\|_2$, where, using (2),

$$\|x - s\|_2 = \|x - Q\|_2 - \|s - Q\|_2 = \tau \sin(\alpha/2) - (\tau - \sigma)\sin(\alpha'/2)$$

and using (5),

$$\|s - u\|_2 \leq \frac{\delta}{2\sin(\alpha'/4)}. \tag{11}$$

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

$$\begin{aligned} r &\geq \frac{\delta}{2} + \|x - s\|_2 + \frac{\delta}{2\sin(\alpha'/4)} \\ &\geq \|y - x\|_2 + \|x - s\|_2 + \|s - u\|_2. \end{aligned} \tag{12}$$

Another condition on r arises when we label edge points far from actual vertices. See Figure 4 (right). If $y' \in \mathbb{Y}$, then it should be labeled as an edge point. That is, at Step 4, the Rips graph $\mathcal{R}_\delta(S_{y'})$ on the shell $S_{y'}$ should have 2 connected components. Therefore the distance $\|F - G\|_2$ between them must be greater than δ. We require that

$$r \geq 2\sigma + \delta/\sqrt{2} \tag{13}$$

which implies $\|F - G\|_2 > \delta$ when r is small enough, as implied by (10).
Note that the value $r = \frac{\delta}{2} + \sigma + \|x - s\|_2 + \frac{\delta}{2\sin(\alpha'/4)}$ satisfies both (12) and (13).

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 $y \in \mathbb{Y}$ in each connected component of the shell, since $\mathbb{Y}$ is $\frac{\delta}{2}$-dense in G_σ.

In the last part of this proof we show that condition (10) is needed to guarantee that the sample is dense enough and the radius of the shells of Step 3 has the correct size, so that, even in the worst case, each vertex is associated to one set of sampled points at Steps 15-17 and these connected components are correctly linked by sets of sampled points labeled as edge points.
Let $z \in G_\sigma$ be the point around e_2 where 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. See Figure 5. If this segment does not exist we simply consider the segment of length $r + \delta$ from s to a point z on e_2.

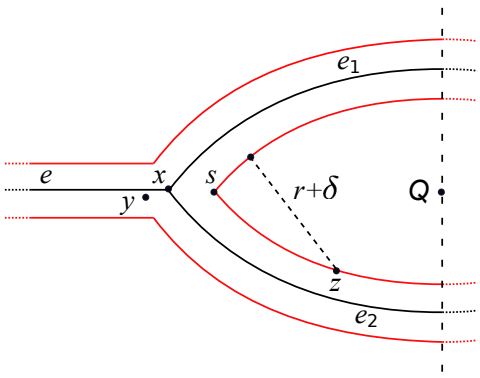

Figure 5: The shell around z is tangent to edge e_2.

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 labeling of the points in the tube between y and z we would like to label all the points in $\{y' \in \mathbb{Y} : \|y' - y\|_2 \leq \|y - z\|_2\}$ as vertices. To simplify the calculation we use the following bound

$$\|y - z\|_2 \leq \|y - x\|_2 + \|x - s\|_2 + \|s - z\|_2,$$

where, using (5),

$$\|s - z\|_2 \leq \frac{r + \delta}{\sin(\alpha'/2)}. \tag{14}$$

This justifies the choice of $p_{11} = \frac{\delta}{2} + \|x - s\|_2 + \frac{r+\delta}{\sin(\alpha'/2)} \geq \|y - z\|_2$. Thus, at Step 11 we label all the points in $\{y' \in \mathbb{Y} : \|y' - y\|_2 \leq p_{11}$ 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.
Finally we determine under which conditions 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 α at both their extremes x and x'. See Figure 6.

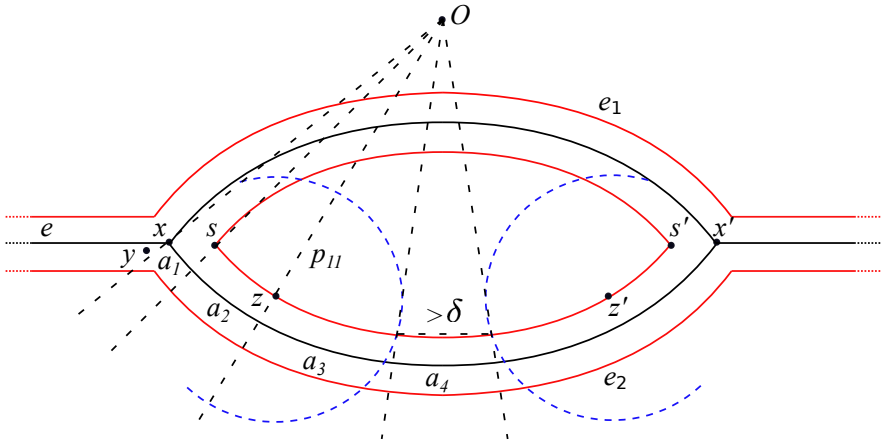

Figure 6: Edges e_1 and e_2, forming an angle of size α at both their extremes x and x'.

All the points $y' \in \mathbb{Y}$ such that $\|y' - z\|_2 \leq p_{11}$ or $\|y' - z'\|_2 \leq p_{11}$ might be labeled as vertices. When we construct $\mathcal{R}(\mathbb{E})_\delta$ and $\mathcal{R}(\mathbb{V})_\delta$ at Step 15 the two sets of vertices around x and x' must be disconnected and there must be at least an edge point between them. 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 $\overline{Ox}$ and $\overline{Os}$ on e_2,
- a_2 is the length of the arc of e_2 formed by the projection of the chord of length $\|s - z\|_2$,
- 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 δ.

Note that, in Figure 6, $e_2 = 2\tau \arcsin\left(\frac{\|x-x'\|_2}{2\tau}\right) = \alpha\tau$ but in general it might be shorter, so that e_1 and e_2 might not intersect in x'. However, by assumptions A2, e_2 must be longer than b. Thus we require

$$\min(b, \alpha\tau) > 2(a_1 + a_2 + a_3) + a_4. \tag{15}$$

By simple properties involving arcs and chords we have

$$a_1 = \left(\frac{\alpha - \alpha'}{2}\right)\tau, \qquad a_2 = 2\tau \arcsin\left(\frac{\|s - z\|_2}{2(\tau - \sigma)}\right),$$
$$a_3 = 2\tau \arcsin\left(\frac{p_{11}}{2(\tau - \sigma)}\right), \quad a_4 = 2\tau \arcsin\left(\frac{\delta}{2(\tau - \sigma)}\right).$$

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

$$\min(b, \alpha\tau) - (\alpha - \alpha')\tau > 2\tau \arcsin\left(\frac{2\|s-z\|_2 + 2p_{11} + \delta}{2(\tau - \sigma)}\right),$$

which holds if

$$\sin\left(\frac{\min(b, \alpha\tau) - (\alpha - \alpha')\tau}{2\tau}\right) > \frac{2\frac{r+\delta}{\sin(\alpha'/2)} + 2p_{11} + \delta}{2(\tau - \sigma)}.$$

The last condition is equivalent to (10). If this condition is satisfied then the graph is correctly reconstructed at Steps 15-17: every connected component of $\mathcal{R}_\delta(\mathbb{V})$ corresponds to a vertex of G and every connected component of $\mathcal{R}_\delta(\mathbb{E})$ corresponds to an edge of G. ■

Example 1 ***A Neuron in Three-Dimensions.*** *We return to the neuron example and we try to apply Propositions 4 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 reach* $30\mu m$ *and* $\xi = 50\mu m$*. The conditions of Proposition 4 are satisfied for* $\delta = 2.00\mu m$*. Algorithm 1 reconstructs the topology of the metric graph starting from a* $\delta/2$*-dense sample. Figure 1b shows the reconstructed graph.*

4. Minimax Analysis

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

$$R_n = \inf_{\widehat{G}} \sup_{P \in \mathcal{P}} P^n\Big(\widehat{G} \not\simeq G\Big), \tag{16}$$

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

4.1 Lower Bounds

To derive a lower bound on the minimax risk, we make repeated use of Le Cam's lemma. See, e.g., Yu (1997) and Chapter 2 of Tsybakov (2008). Recall that the total variation distance between two measures P and Q on the same probability space is defined by $\mathrm{TV}(P,Q) = \sup_A |P(A) - Q(A)|$ where the supremum is over all measurable sets. It can be shown that $\mathrm{TV}(P,Q) = P(H) - Q(H)$, where $H = \{y : p(y) \geq q(y)\}$ and p and q are the densities of P and Q with respect to any measure that dominates both P and Q.

Lemma 5 (Le Cam) *Let $\mathcal{Q}$ be a set of distributions. Let $\theta(Q)$ take values in a metric space with metric ρ. 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*

$$\inf_{\widehat{\theta}} \sup_{Q \in \mathcal{Q}} \mathbb{E}_{Q^n}\Big[\rho(\widehat{\theta}, \theta(Q))\Big] \geq \frac{1}{8}\rho(\theta(Q_1), \theta(Q_2))(1 - TV(Q_1, Q_2))^{2n} \tag{17}$$

where the infimum is over all the estimators of $\theta(Q)$.

Below we apply Le Cam's lemma using several pairs of distributions. Any pair Q_1, Q_2 is associated with a pair of metric graphs $G', G'' \in \mathcal{G}$. We take $\theta(Q_1)$ and $\theta(Q_2)$ to be the classes of graphs that are isomorphic to G' and G''. We set $\rho(\theta(Q_1), \theta(Q_2)) = 0$ if G' and G'' are isomorphic and $\rho(\theta(Q_1), \theta(Q_2)) = 1$ otherwise. Figure 7 shows several pairs of metric graphs that are used to derive lower bounds in the noiseless case and in the tubular noise case. In the noiseless case we ignore the σ-tubes around the metric graphs.

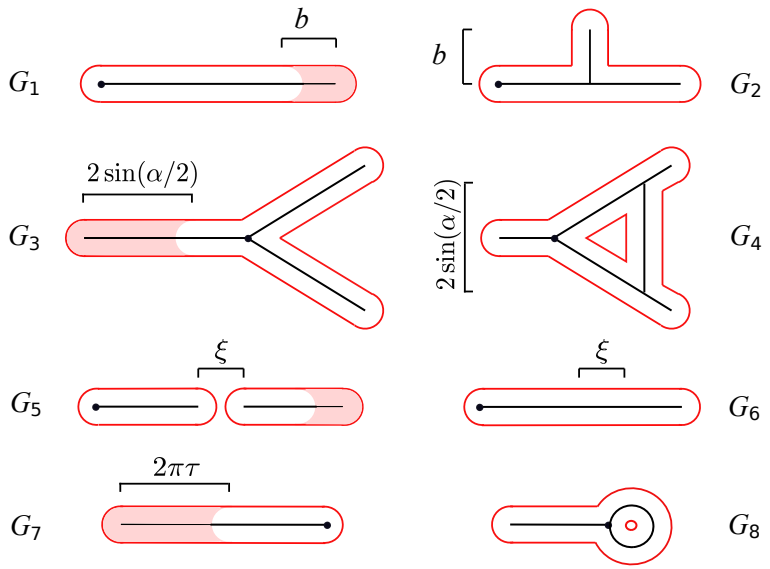

Figure 7: Pairs of metric graphs used in the derivation of lower bounds in the noiseless case and in the tubular noise case.

Recall that, in the noiseless case, we restrict the attention to 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 $a > 0$.

Theorem 6 *In the noiseless case ($\sigma = 0$), 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*

$$R_n \geq \exp\Big(-2a \min\{b, 2\sin(\alpha/2), \xi, 2\pi\tau\} n\Big). \tag{18}$$

Proof We consider the 4 parameters separately. See Figure 7, ignoring the red lines representing the tubular noise that is not considered in this theorem.

Shortest edge b. Consider the metric graph G_1 consisting of a single edge of length $1+b$ and metric graph G_2 with an edge of length 1 and an orthogonal edge of length b glued in the middle. The density on G_1 is constructed in the following way: on the set $G_1 \backslash G_2$ 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 $G_2 \backslash G_1$, which correspond to the orthogonal edge of length b. We evenly spread the remaining mass. The two densities differ

only on the sets $G_1 \backslash G_2$ and $G_2 \backslash G_1$. Therefore $\mathrm{TV}(p_1, p_2) \leq ab$ and, by Le Cam's lemma, $R_n \geq \frac{1}{8}(1-ab)^{2n} \geq \frac{1}{8}\mathrm{e}^{-2abn}$ for all $b \leq b_0(a)$, where $b_0(a)$ is a constant depending on a.

Smallest angle α. Now consider the metric graphs G_3 and G_4. G_3 consists of two edges of length 2 forming an angle α 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 $G_3 \backslash G_4$, $p_4(x) = a$ on $G_4 \backslash G_3$ and spread evenly the rest of the mass. The total variation distance is $\mathrm{TV}(p_3, p_4) \leq 2a \sin\left(\frac{\alpha}{2}\right)$ and, by Le Cam's lemma, $R_n \geq \frac{1}{8}(1 - 2a\sin(\alpha/2))^{2n} \geq \frac{1}{8}\mathrm{e}^{-4a\sin(\alpha/2)n}$ for all $\alpha \leq \alpha_0(a)$, where $\alpha_0(a)$ is a constant depending on a.

Global reach ξ. We defined the global reach as the shortest Euclidean distance between two points that are far apart in the graph distance. Figure 7 shows 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, ξ apart from each other. Again, we set $p_5(x) = a$ on $G_5 \backslash G_6$, $p_6(x) = a$ on $G_6 \backslash G_5$ and evenly spread the rest. We obtain $\mathrm{TV}(p_5, p_6) \leq a\xi$ and, by Le Cam's lemma, $R_n \geq \frac{1}{8}(1-a\xi)^{2n} \geq \frac{1}{8}\mathrm{e}^{-2a\xi n}$ for all $\xi \leq \xi_0(a)$, where $\xi_0(a)$ is a constant depending on a.

Local reach τ. The local reach τ is the smallest reach of the edges forming the metric graph. Consider metric graphs G_7 and G_8. G_7 consists of a loop of radius τ attached to an edge of length 1 and metric graph G_8 is a single edge of length $1 + 2\pi\tau$. As in the previous cases $p_7(x) = a$ on $G_7 \backslash G_8$ and $p_8(x) = a$ on $G_8 \backslash G_7$. It follows that $\mathrm{TV}(p_7, p_8) \leq 2a\pi\tau$ and, by Le Cam's lemma, $R_n \geq \frac{1}{8}(1 - 2a\pi\tau)^{2n} \geq \frac{1}{8}\mathrm{e}^{-4a\pi\tau n}$ for all $\tau \leq \tau_0(a)$, where $\tau_0(a)$ is a constant depending on a. ■

For the tubular noise case we assume that σ is small enough to guarantee that $R_n < 1$, that is, the problem is not hopeless. In particular, we require that σ satisfies conditions (9) and (10) of Proposition 4, which can be combined into the following condition

$$0 < \min\left\{ \frac{\xi - 3\sigma - \tau\sin(\alpha/2) + (\tau - \sigma)\sin(\alpha'/2)}{3/2 + [2\sin(\alpha'/4)]^{-1}},\ f(b, \alpha, \tau, \xi, \sigma) \right\}. \tag{19}$$

Theorem 7 *Assume that σ is positive and satisfies condition* (19)*. In the tubular noise case, for $b \leq b_0(D)$, $\alpha \leq \alpha_0(D)$, $\xi \leq \xi_0(D)$, $\tau \leq \tau_0(D)$, where $b_0(D), \alpha_0(D), \xi_0(D)$ and $\tau_0(D)$ are constants which depend on the ambient dimension D, a lower bound on the minimax risk for metric graph reconstruction is*

$$R_n \geq \frac{1}{8}\exp\Big(-2\min\{C_{D,1}b, C_{D,2}\sin(\alpha/2), C_{D,3}\xi, C_{D,4}\tau\}n\Big), \tag{20}$$

for some constants $C_{D,1}, C_{D,2}, C_{D,3}, C_{D,4}$.

Proof As in the proof oh Theorem 6 we consider the 4 parameters separately. We compare the pairs of graphs shown in Figure 7, including the tubular regions constructed around them, from which we get samples uniformly.

Shortest edge b. Consider the metric graph G_1 consisting of a single edge of length $1+b$ and metric graph G_2 with an edge of length 1 and an orthogonal edge of length b glued in the middle. Since $\mathrm{vol}(G_1) > \mathrm{vol}(G_2)$, the density q_1 at a point in the tube around G_1 is lower than the density q_2 at a point around G_2. From the definition of total variation

$TV = q_1(H) - q_2(H)$ where H is the set where $q_1 > q_2$, the shaded area in Figure 7. Note that $q_2(H) = 0$ and

$$TV(q_1, q_2) = q_1(H) = \frac{\text{vol}(H)}{\text{vol}(G_1)} \leq C_{D,1} \frac{b\sigma^{D-1}}{(1+b)\sigma^{D-1}} \leq C_{D,1} b.$$

By Le Cam's lemma, $R_n \geq \frac{1}{8}(1 - C_{D,1}b)^{2n} \geq \frac{1}{8}\mathrm{e}^{-2C_{D,1}bn}$ for all $b \leq b_0(D)$, where $b_0(D)$ is a constant depending on D.

Smallest angle α. Now consider the metric graphs G_3 and G_4. Since $\text{vol}(G_3) > \text{vol}(G_4)$, the density q_3 at a point in the tube around G_3 is lower than the density q_4 at a point around G_4. $TV = q_3(H) - q_4(H)$ where H is the set where $q_3 > q_4$, the shaded area in the tube around G_3. Note that $q_4(H) = 0$ and

$$TV(q_3, q_4) = q_3(H) = \frac{\text{vol}(H)}{\text{vol}(G_3)} \leq C_{D,2} \frac{\sin(\alpha/2)\sigma^{D-1}}{(1+\sin(\alpha/2))\sigma^{D-1}} \leq C_{D,2} \sin(\alpha/2).$$

By Le Cam's lemma, $R_n \geq \frac{1}{8}(1 - C_{D,2}\sin(\alpha/2))^{2n} \geq \frac{1}{8}\mathrm{e}^{-2C_{D,2}\sin(\alpha/2)n}$ for all $\alpha \leq \alpha_0(D)$, where $\alpha_0(D)$ is a constant depending on D.

Global reach ξ. Figure 7 shows 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, ξ apart from each other. Since $\text{vol}(G_5) > \text{vol}(G_6)$, the density q_5 at a point in the tube around G_5 is lower than the density q_6 at a point around G_6. $TV = q_5(H) - q_6(H)$ where H is the set where $q_5 > q_6$, the shaded area in the tube around G_5. Note that $q_6(H) = 0$ and

$$TV(q_5, q_6) = q_5(H) = \frac{\text{vol}(H)}{\text{vol}(G_5)} \leq C_{D,3} \frac{\xi\sigma^{D-1}}{\sigma^{D-1}} = C_{D,3}\xi.$$

By Le Cam's lemma, $R_n \geq \frac{1}{8}(1 - C_{D,3}\xi)^{2n} \geq \frac{1}{8}\mathrm{e}^{-2C_{D,3}\xi n}$ for all $\xi \leq \xi_0(D)$, where $\xi_0(D)$ is a constant depending on D.

Local reach τ. The local reach τ is the smallest reach of the edges forming the metric graph. Consider metric graphs G_7 and G_8 in Figure 7. Since $\text{vol}(G_7) > \text{vol}(G_8)$, the density q_7 at a point in the tube around G_7 is lower than the density q_8 at a point around G_8. $TV = q_7(H) - q_8(H)$ where H is the set where $q_7 > q_8$, the shaded area in the tube around G_7. Note that $q_8(H) = 0$ and

$$TV(q_7, q_8) = q_7(H) = \frac{\text{vol}(H)}{\text{vol}(G_7)} \leq C_{D,4} \frac{\tau\sigma^{D-1}}{(1+\tau)\sigma^{D-1}} \leq C_{D,4}\tau.$$

By Le Cam's lemma, $R_n \geq \frac{1}{8}(1 - C_{D,4}\tau)^{2n} \geq \frac{1}{8}\mathrm{e}^{-2C_{D,4}\tau n}$ for all $\tau \leq \tau_0(D)$, where $\xi_0(D)$ is a constant depending on D. ■

Note that, up to constants, the lower bound obtained in the tubular noise case is identical to the lower bound of Proposition 6 for the noiseless case.

4.2 Upper Bounds

In this section we use the analysis of the performance of Algorithm 1 to derive an upper bound on the minimax risk. We will use the strategy of Niyogi et al. (2008) to find the

sample size that guarantees a $\delta/2$-dense sample with high probability. We will use the following two lemmas.

Lemma 8 (5.1 in Niyogi et al. 2008) *Let $\{A_i\}$ for $i = 1, \ldots, l$ be a finite collection of measurable sets and let μ 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 μ. 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 ϵ-covering number $C(\epsilon)$ of a set S is the smallest number of Euclidean balls of radius ϵ required to cover the set. The ϵ-packing number $P(\epsilon)$ is the maximum number of sets of the form $B(x, \epsilon) \cap S$, where $x \in S$, that may be packed into S without overlap.

Lemma 9 (5.2 in Niyogi et al. 2008) *For every $\epsilon > 0$, $P(2\epsilon) \leq C(2\epsilon) \leq P(\epsilon)$.*

Combining Lemma 8 and Proposition 4, we obtain an upper bound on R_n for the noiseless case.

Theorem 10 *In the noiseless case ($\sigma = 0$), an upper bound on the minimax risk R_n is given by*

$$R_n \leq \frac{8 \; \mathit{length}(G)}{\delta} \; \exp\left\{-\frac{a\,\delta\,n}{4 \; \mathit{length}(G)}\right\},$$

where

$$\delta = \frac{1}{2}\min\left\{\xi\frac{2\sin(\alpha/4)}{3\sin(\alpha/4)+1} \;,\; \frac{\tau\sin(\alpha/2)\sin(\alpha/4)}{\sin(\alpha/2)\sin(\alpha/4)+3\sin(\alpha/4)+1}\sin\left(\frac{\min\{b, \alpha\tau\}}{2\tau}\right)\right\}. \tag{21}$$

Proof In the noiseless case, Proposition 4 implies that the graph G can be reconstructed from a $\delta/2$-dense sample $\mathbb{Y}$ if

$$\delta < \min\left\{\xi\frac{2\sin(\alpha/4)}{3\sin(\alpha/4)+1} \;,\; f(b, \alpha, \tau, \xi, 0)\right\}. \tag{22}$$

The value of δ selected in (21) satisfies condition (22), which follows from conditions (9) and (10), with $\sigma = 0$. 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 $A_i^{\delta/4} = B_{\delta/4}(x_i) \cap G$. Applying Lemma 8 we find that the sample size that guarantees a correct reconstruction with probability at least $1 - \lambda$ is

$$\frac{1}{\gamma}\left(\log l + \log\frac{1}{\lambda}\right), \tag{23}$$

where

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

and we bound the covering number l in terms of the packing number, using Lemma 9:

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

Therefore, from (23), if

$$n = \frac{4 \text{ length } (G)}{a\delta} \left[\log \left(\frac{8 \text{ length}(G)}{\delta} \right) + \log \frac{1}{\lambda} \right] \tag{24}$$

we have a $\delta/2$-dense sample with probability at least $1 - \lambda$ and, by Proposition 4, $\mathbb{P}(\widehat{G} \not\approx G) \leq \lambda$. Rearranging we have the result. ■

Note that, in the noiseless case, the upper and lower bounds are tight up to polynomial factors in the parameters τ, b, ξ. There is a small gap with respect to α; closing this gap is an open problem.
In the tubular noise case, we assume that σ is small enough, to guarantee that Algorithm 1 correctly reconstructs a metric graph starting from a $\delta/2$-dense sample.

Theorem 11 *Assume that σ satisfies condition* (19) *and $0 < \sigma < \min\{3\tau/16, \delta/8\}$, where*

$$\delta = C_0 \min \left\{ \frac{\xi - 3\sigma - \tau \sin(\alpha/2) - (\tau - \sigma)\sin(\alpha'/2)}{3/2 + [2\sin(\alpha'/4)]^{-1}},\ f(b, \alpha, \tau, \xi, \sigma) \right\}, \tag{25}$$

for some $0 < C_0 < 1$. Under the tubular noise model, an upper bound on the minimax risk R_n is given by

$$R_n \leq \frac{16 length(G)}{\delta} \exp\left(-\frac{C'_D \delta(\tau - 8\sigma)n}{\tau \, length(G)} \right),$$

where C'_D is a constant depending on the ambient dimension.

Proof Proposition 4 implies that the graph G can be reconstructed from a $\delta/2$-dense sample $\mathbb{Y}$ if

$$\delta < \min \left\{ \frac{\xi - 3\sigma - \tau \sin(\alpha/2) - (\tau - \sigma)\sin(\alpha'/2)}{3/2 + [2\sin(\alpha'/4)]^{-1}},\ f(b, \alpha, \tau, \xi, \sigma) \right\}, \tag{26}$$

which is satisfied by the value of δ selected in (25). We look for the sample size n that guarantees a $\delta/2$-dense sample in G_σ with high probability.
We consider a cover of the metric graph G by Euclidean balls of radius $\delta/8$. Let $\{x_i : 1 \leq i \leq l\}$ be the centers of such balls that constitute a minimal cover. Note that D-dimensional balls of radius $\delta/8 + \sigma \leq \delta/4$ centered at the same $x_i's$ constitute a cover of the tubular region G_σ. We define $A_i^{\delta/8+\sigma} = B_{\delta/8+\sigma}(x_i) \cap G_\sigma$. Applying Lemma 8 we find that the sample size

that guarantees a $\delta/2$-dense sample in G_σ (and a correct topological reconstruction of G) with probability at least $1-\lambda$ is

$$\frac{1}{\gamma}\left(\log l + \log\frac{1}{\lambda}\right), \tag{27}$$

where

$$\gamma = \min_i \frac{\mathrm{vol}(A_i^{\delta/8+\sigma})}{\mathrm{vol}(G_\sigma)}. \tag{28}$$

Define $\tilde{A}_i^\delta = B_\delta(x_i)\cap G$. The covering number l is bounded in terms of the packing number, using Lemma 9,

$$l \leq \frac{\mathrm{length}(G)}{\min_i \mathrm{length}(\tilde{A}_i^{\delta/16})} \leq \frac{16\ \mathrm{length}(G)}{\delta}.$$

We construct a lower bound on γ by deriving an upper bound on the denominator of (28) and a lower bound on the numerator.

Upper bound on vol(G_σ). Let N_σ be the σ-covering number of G and let $\mathcal{C}_\sigma$ be the set of centers of this cover. By Lemma 9, N_σ is bounded by the $\sigma/2$-packing number. A simple volume argument gives $N_\sigma \leq C\mathrm{length}(G)/\sigma$, for some constant C. Note that 2σ D-dimensional balls around each of the centers in $\mathcal{C}_\sigma$ cover G_σ. Thus $\mathrm{vol}(G_\sigma) \leq v_D N_\sigma (2\sigma)^D \leq C_D \mathrm{length}(G)\sigma^{D-1}$ for some constant C_D depending on the ambient dimension.

Lower bound on vol$(A_i^{\delta/8+\sigma})$, for all i. Let $P_A(\sigma)$ be the σ-packing number of $\tilde{A}_i^{\delta/8}$ and let $\mathcal{D}_A$ be the set of centers of this packing. Then $\mathrm{vol}(A_i^{\delta/8+\sigma}) \geq P_A(\sigma) v_D \sigma^D$, because the union of σ balls around $\mathcal{D}_A$ is contained in $A_i^{\delta/8+\sigma}$. Let $C_A(2\sigma)$ be the 2σ-covering number of $\tilde{A}_i^{\delta/8}$ and let $\mathcal{C}_A = \{z_1, \ldots, z_{C_A(2\sigma)}\}$ be the set of centers of this cover. By Lemma 9,

$$P_A(\sigma) \geq C_A(2\sigma) \geq \frac{\mathrm{length}(\tilde{A}_i^{\delta/8})}{\max_{z_j\in\mathcal{C}_A} \mathrm{length}(B_{2\sigma}(z_j)\cap \tilde{A}_i^{\delta/8})} \geq \frac{\delta/8}{\max_{z_j\in\mathcal{C}_A} \mathrm{length}(B_{2\sigma}(z_j)\cap \tilde{A}_i^{\delta/8})}$$

and, since $2\sigma < 3\tau/8$, by Corollary 1.3 in Chazal (2013),

$$\max_{z_j\in\mathcal{C}_A} \mathrm{length}(B_{2\sigma}(z_j)\cap \tilde{A}_i^{\delta/8}) \leq C_2\left(\frac{\tau}{\tau-8\sigma}\right)\sigma,$$

for some constant C_2. Thus

$$\gamma \geq \frac{P_A(\sigma) v_D \sigma^D}{C_D \mathrm{length}(G)\sigma^{D-1}} \geq C'_D \frac{\delta(\tau-8\sigma)}{\tau \mathrm{length}(G)} \quad ,$$

where C'_D is a constant depending on the ambient dimension.

Finally, from (27), if

$$n = \frac{\tau\ \mathrm{length}\ (G)}{C'_D\delta(\tau-8\sigma)}\left[\log\left(\frac{16\ \mathrm{length}(G)}{\delta}\right) + \log\frac{1}{\lambda}\right], \tag{29}$$

then the sample is $\delta/2$-dense with probability at least $1-\lambda$ and $\mathbb{P}(\widehat{G} \not\simeq G) \leq \lambda$. Rearranging we obtain

$$R_n \leq \exp\left(-\frac{C'_D \delta(\tau - 8\sigma)n}{\tau\, \text{length}(G)} + \log\left(\frac{16\text{length}(G)}{\delta}\right)\right).$$

■

5. Discussion

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. In Proposition 4 we have analyzed Algorithm 1 using the Euclidean distance at every step. It is possible to obtain a similar result using a different notion of distance, for example, the distance induced by a Rips-Vietoris graph constructed on the sample.

While in our analysis we mainly relied on the assumption of a dense sample, Aanjaneya et al. (2012) used the more refined but stronger assumption of the sample being an approximation of the metric graph, which we recall: given positive numbers ε and R, we say that $(\mathbb{Y}, d_{\mathbb{Y}})$ is an (ε, R)-*approximation* of the metric space (G, d_G) if there exists a correspondence $C \subset G \times \mathbb{Y}$ such that

$$(x,y),(x',y') \in C, \min(d_G(x,x'), d_{\mathbb{Y}}(y,y')) \leq R \implies \left|d_G(x,x') - d_{\mathbb{Y}}(y,y')\right| \leq \varepsilon. \tag{30}$$

As shown in Aanjaneya et al. (2012), the (ε, R)-approximation assumption is sufficient, for appropriate choice of the parameters ε 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. A strategy similar to the one used in this paper could be used to determine the sample size that guarantees an (ε, R)-approximation of the underlying metric graph with high probability. This would guarantee a correct topological reconstruction, as well as an approximation of the metric d_G.

We are also 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., 2012b) 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

Research supported by NSF CAREER Grant DMS 114967, Air Force Grant FA95500910373, NSF Grant DMS-0806009. The authors thank the referees for helpful comments and suggestions.

References

Mridul Aanjaneya, Frédéric Chazal, Daniel Chen, Marc Glisse, Leonidas Guibas, and Dmitriy Morozov. Metric graph reconstruction from noisy data. *International Journal of Computational Geometry & 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, 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.

Frédéric Chazal. An upper bound for the volume of geodesic balls in submanifolds of Euclidean spaces. Technical report, INRIA, January 2013.

Frédéric Chazal and André 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.

Frédéric Chazal and Jian Sun. Gromov-Hausdorff approximation of metric spaces with linear structure. *arXiv preprint arXiv:1305.1172*, 2013.

Frédéric Chazal, David Cohen-Steiner, and André Lieutier. A sampling theory for compact sets in Euclidean space. *Discrete & Computational Geometry*, 41(3):461–479, 2009.

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. SIAM, 2010.

Alexey Chernov and Vitaliy Kurlin. Reconstructing persistent graph structures from noisy images. *Electronic Journal Image-A*, 3(5):19–22, 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.

Herbert Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93(3):418–491, 1959.

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.

Xiaoyin Ge, Issam I Safa, Mikhail Belkin, and Yusu Wang. Data skeletonization via Reeb graphs. In *Advances in Neural Information Processing Systems*, pages 837–845, 2011.

Christopher R Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Minimax manifold estimation. *Journal of Machine Learning Research*, 13:1263–1291, 2012a.

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

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.

Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39(1-3):419–441, 2008.

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2008.

Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.