Finding the Homology of Submanifolds with High Confidence from Random Samples

P. Niyogi[*], S. Smale[†], S. Weinberger[‡]

September 13, 2004

Abstract

Recently there has been a lot of interest in geometrically motivated approaches to data analysis in high dimensional spaces. We consider the case where data is drawn from sampling a probability distribution that has support on or near a submanifold of Euclidean space. We show how to "learn" the homology of the submanifold with high confidence. We discuss an algorithm to do this and provide learning-theoretic complexity bounds. Our bounds are obtained in terms of a condition number that limits the curvature and nearness to self-intersection of the submanifold. We are also able to treat the situation where the data is "noisy" and lies near rather than on the submanifold in question.

1 Introduction

In recent years, there has been considerable interest in the possibility of analyzing and processing data in high dimensional spaces. Following the intuition that naturally occurring data may be generated by structured

[*] Departments of Computer Science, Statistics, University of Chicago.

[†] Toyota Technological Institute, Chicago.

[‡] Department of Mathematics, University of Chicago.

systems with possibly much fewer degrees of freedom than the ambient dimension would suggest, various researchers (see [5, 2, 3, 6, 1] have considered the case when the data lives on or close to a submanifold of the ambient space. One hopes then to estimate geometrical and topological properties of the submanifold from random points ("scattered data") lying on this unknown submanifold.

In this paper, we consider the particular question of identifying the homology of the submanifold from random samples. The homology of the submanifold (see [9] for definitions) are natural topological invariants that provide a good characterization of many aspects of it. For example, the dimensions of the homology groups, the Betti numbers ($\beta_0, \beta_1 \ldots$) have natural interpretations. β_0, the dimension of the zeroth homology group is the number of connected components of the submanifold. In data analysis situations, the number of clusters of the data may sometimes be understood in terms of the number of components of an underlying manifold (or other geometric object). If the dimension of the submanifold is d, then one sees that $\beta_j = 0$ for all $j > d$. Thus the the largest non-trivial homology gives us the dimension of the submanifold. If the submanifold is two-dimensional, then β_0 and β_1, are related to the number of connected components and number of holes respectively of the submanifold.

We show that it is possible to identify the homology from random samples and discuss an algorithm to do this. There are a few aspects of the developments in this paper that are worth emphasizing. First, we provide sample complexity estimates on the number of examples that are needed to identify the homology with high confidence. Our results are in the style of learning theoretic treatments where unknown objects (typically functions in learning theory) are "learned" from random samples and confidence estimates are provided. Second, we treat the situation where data might be drawn from a distribution that is concentrated *around* the manifold rather than precisely on it. Under specific models of noise, we show that our algorithm can work even with noisy data. In all cases, estimates are provided in terms of a condition number that limits the curvature and nearness to self-intersection of the submanifold.

Our results may also be of interest to researchers in computational geometry and topology who have considered the question of computing homology from simplicial complexes in the past (see [13, 7] for details and further references). Researchers in graphics, pattern recognition, solid modeling, molecular biology, finance, and other areas where large amounts of high dimensional data are available may find some use for the topological perspective on data analysis embodied in the algorithms and analyses of this paper.

2 Preliminaries

Consider a compact Riemannian submanifold $\mathcal{M}$ of a Euclidean space $\mathbb{R}^k$. Sample the manifold according to a uniform probability measure on it. Thus points $x_1, \ldots, x_n \in \mathcal{M}$ are generated. This set of points $\bar{x} = \{x_1, \ldots, x_n\}$ will be the data set on the basis of which homology groups will be calculated. In later sections, we will consider the case when the data is drawn from a probability measure with support close to the manifold.

Throughout our discussion, we will associate to $\mathcal{M}$ a condition number $(1/\tau)$ where τ defined as the largest number having the property: The open normal bundle about $\mathcal{M}$ of radius r is imbedded in $\mathbb{R}^k$ for every $r < \tau$. Its image Tub_τ is a tubular neighborhood of $\mathcal{M}$ with its canonical projection map

$$\pi_0 : \text{Tub}_\tau \to \mathcal{M}$$

Note that τ encodes both local curvature considerations as well as global ones: If $\mathcal{M}$ is a union of several components τ bounds their separation. For example, if $\mathcal{M}$ is a sphere, then τ is equal to its radius. If $\mathcal{M}$ is an annulus, then τ is the separation of its components. In Section 6 we relate the condition number $\frac{1}{\tau}$ to classical notions of curvature in differential geometry via the second fundamental form.

3 An Outline of our Main Results

Ultimately we wish to compute the homology of the manifold $\mathcal{M} \subset \mathbb{R}^k$ from the randomly sampled datapoints $\bar{x} = \{x_1, \ldots, x_n\} \subset \mathcal{M}$. We first begin by considering Euclidean balls (in the ambient space $\mathbb{R}^k$) of radius ϵ and centers x_i's. Let us denote these balls as $B_\epsilon(x_i)$. We can now define the open set $U \subset \mathbb{R}^k$ given by

$$U = \cup_{x \in \bar{x}} B_\epsilon(x)$$

Our first proposition states that if $\bar{x} = \{x_1, \ldots, x_n\}$ is $\epsilon/2$ dense in $\mathcal{M}$, then $\mathcal{M}$ is a deformation retract of U.

Proposition 3.1 *Let $\bar{x}$ be any finite collection of points $x_1, \ldots, x_n \in \mathrm{IR}^k$ such that it is $\frac{\epsilon}{2}$ dense in $\mathcal{M}$, i.e., for every $p \in \mathcal{M}$, there exists an $x \in \bar{x}$ such that $\| p - x \|_{\mathrm{IR}^k} < \frac{\epsilon}{2}$. Then for any $\epsilon < \sqrt{\frac{3}{5}}\tau$, we have that U deformation retracts to $\mathcal{M}$. Therefore homology of U equals homology of $\mathcal{M}$.*

We prove this proposition in Section 4.
In the case under consideration here, the points $x_1, \ldots, x_n$ are sampled in i.i.d. fashion from the uniform probability distribution on $\mathcal{M}$. By probabilistic considerations, we will then prove (in Section 5)

Proposition 3.2 *Let $\bar{x}$ be drawn by sampling $\mathcal{M}$ in i.i.d. fashion according to the uniform probability measure on $\mathcal{M}$. Then with probability greater than $1 - \delta$, we have that $\bar{x}$ is $\frac{\epsilon}{2}$-dense ($\epsilon < \frac{\tau}{2}$) in $\mathcal{M}$ provided*

$$|\bar{x}| > \beta_1(\log(\beta_2) + \log(\frac{1}{\delta}))$$

where $\beta_1 = \frac{vol(\mathcal{M})}{(\cos(\theta))^k vol(B^k_\epsilon)}$ and $\beta_2 = \frac{vol(\mathcal{M})}{(\cos(\theta))^k vol(B^k_{\epsilon/8})}$. Here $vol(B^k_\epsilon)$ denotes the k-dimensional volume of the standard k-dimensional ball of radius ϵ. Finally $\theta = \arcsin(\frac{\epsilon}{2\tau})$.

Putting these two propositions together, we see that we are able to provide a finite sample estimate for how many times we need to sample $\mathcal{M}$ so that we are guaranteed with high confidence that the homology of the random set U equals the homology of $\mathcal{M}$. Thus our main theorem is

Theorem 3.1 *Let $\mathcal{M}$ be a compact submanifold of IR^k with condition number τ. Let $\bar{x} = \{x_1, \ldots, x_n\}$ be a set of n points drawn in i.i.d fashion according to the uniform probability measure on $\mathcal{M}$. Let $0 < \epsilon < \frac{\tau}{2}$. Let $U = \cup_{x \in \bar{x}} B_\epsilon(x)$ be a correspondingly random open subset of IR^k. Then for all*

$$n > \beta_1(\log(\beta_2) + \log(\frac{1}{\delta}))$$

the homology of U equals the homology of $\mathcal{M}$ with high confidence (probability $> 1 - \delta$).

3.1 Computing the Homology of U

One now needs to consider algorithms to compute the homology of U. Noting that the $B_\epsilon(x_i)$'s form a cover of U, one can construct the *nerve* of the cover. The nerve is an abstract simplicial complex constructed as follows: One puts in a k-simplex for every $k+1$-tuple of intersecting elements of the cover. The Nerve Lemma (see [4]) applies in our case, as balls are convex, to show that the homology of U is the same as the homology of this complex. The algorithm consists of the following components.

1. Given an ϵ, and a set of points $\bar{x} = \{x_1, \ldots, x_n\}$ in $\mathbb{R}^k$, each j-simplex is given by a subset of the n points that have non-zero intersection. Thus we may define K_j to be a simplicial complex (of dimension j). Each simplex $\sigma \in K_j$ is associated with a set of $j+1$ points $(p_0(\sigma), \ldots, p_j(\sigma) \in \bar{x})$ such that

   $$\cap_{i=0}^{j} B_\epsilon(p_i(\sigma)) \neq \phi$$

   An orientation for the simplex is chosen by picking an ordering and let us denote the oriented simplex by $|p_0(\sigma), \ldots, p_j(\sigma)|$.

2. A very crude upper bound on the size of K_j (denoted by $|K_j|$) is given by $\binom{n}{j+1}$. However, it is clear that if two points x_m and x_l are more than 2ϵ apart, they cannot be associated to a simplex. Therefore, there is a locality condition that the $p_i(\sigma)$'s must obey which results in $|K_j|$ being much smaller than this crude number. The simplicial complex $K = \cup_{j=0}^{k} K_j$ together with face relations.

3. A basic subroutine for computing the simplicial complex (steps 1 and 2 above) involves the decision problem: for any set of j points, determine whether balls of radius ϵ around each of these points have non-empty intersection. This problem is related to the smallest ball problem defined as follows: Given a set of j points, find the the ball with smallest radius enclosing all these points. One can check that $\cap_{i=1}^{j} B_\epsilon(p_i) \neq \phi$ if and only if this smallest radius $< \epsilon$. Fast algorithms for the smallest ball problem exist. See [11] for theoretical discussion and "http://www2.inf.ethz.ch/personal/gaertner/miniball.html" for downloadable algorithms from the web.

4. A j-chain is a function $c : K_j \to \mathbb{R}$ and can be written as a formal sum

   $$c = \sum_{\sigma \in K_j} c(\sigma)\sigma$$

   By addition j-chains component wise, one gets the vector space of j-chains denoted by C_j.

5. The boundary operator ∂_j is a linear operator from C_j to C_{j-1} defined as follows. For each (oriented) simplex $\sigma \in K_j$,

   $$\partial_j \sigma = \sum_{i=0}^{j} (-1)^i \sigma_i$$

   where σ_i is a $j-1$ face of σ (facing point $p_i(\sigma)$) and the orientation of σ_i is given by $|p_0, \ldots, p_{i-1}, p_{i+1}, \ldots, p_j|$. Now ∂_j is defined on j chains by additivity as

   $$\partial_j (\sum_{\sigma \in K_j} c(\sigma)\sigma) = \sum_{\sigma \in K_j} c(\sigma)\partial_j \sigma$$

   Thus, ∂_j can be represented as a $n_{j-1} \times n_j$ matrix where $n_{j-1} = |K_{j-1}|$ and $n_j = |K_j|$ respectively. The matrix is usually sparse in our setting.

6. This defines the chain complex

   $$\ldots C_{j+1} \rightarrow_{\partial_{j+1}} C_j \rightarrow_{\partial_j} C_{j-1} \ldots$$

   One can finally define the *image* and *kernel* of the boundary operator given by

   $$Im\partial_j = \{c \in C_{j-1} | \exists c' \in C_j \text{ where } \partial_j c' = c\}$$

   and

   $$Ker\partial_j = \{c \in C_j | \partial_j c = 0\}$$

   Now $Im\partial_{j+1}$ is the vector space of j-boundaries and $Ker\partial_j$ is the vector space of j cycles. Then the jth homology group is the quotient of $Ker\partial_j$ over $Im\partial_{j+1}$, i.e.,

   $$H_j = Ker\partial_j \setminus Im\partial_{j+1}$$

The calculation of H_j is seen to be an exercise in linear algebra given the matrix representation of the boundary operators. In our exposition here, we have been working over a field resulting in vector spaces which are characterized purely by their ranks (the Betti numbers). One approach to this is also via the combinatorial Laplacian as outlined in Friedman (1998). More generally, one can work over a module and H_j would then be an Abelian group.

4 The Deformation Retract Argument

In this section we prove Proposition 3.1 Consider the canonical map $\pi : U \rightarrow \mathcal{M}$ given by (π is the restriction of π_0 to U)

$$\pi(x) = \arg\min_{p \in \mathcal{M}} ||x - p||$$

Then we see that the fibers $\pi^{-1}(p)$ are given by

$$\pi^{-1}(p) = \cup_{x \in \bar{x}} B_\epsilon(x) \cap T_p^\perp \cap B_\tau(p)$$

where $T_p^\perp$ is the normal subspace at $p \in \mathcal{M}$ orthogonal to the tangent space T_p. Let us also define $st(p)$ as

$$st(p) = \cup_{\{x \in \bar{x}; x \in B_\epsilon(p)\}} B_\epsilon(x) \cap T_p^\perp \cap B_\tau(p)$$

It is immediately clear that

$$st(p) \subseteq \pi^{-1}(p)$$

Then the following simple proposition is true.

Proposition 4.1 *$st(p)$ is star shaped and therefore contracts to p.*

PROOF:Consider arbitrary $v \in st(p)$. Then $v \in B_\epsilon(x) \cap T_p^\perp$ for some $x \in \bar{x}$ such that $x \in B_\epsilon(p)$. Since $x \in B_\epsilon(p)$, we immediately have $p \in B_\epsilon(x)$. Since v, p are both in $B_\epsilon(x)$, by convexity of Euclidean balls, we have that the line segment $\bar{vp}$ joining v to p is entirely contained in $B_\epsilon(x)$. At the same time, $\bar{vp}$ is entirely contained in $T_p^\perp$ and it follows therefore that $\bar{vp}$ is contained in $st(p)$.

□

We next show that the inclusion of $st(p)$ in $\pi^{-1}(p)$ is an equality proving that $\pi^{-1}(p)$ contracts to p.

Proposition 4.2

$$st(p) = \pi^{-1}(p)$$

PROOF:We need to show that $\pi^{-1}(p) \subseteq st(p)$. Consider an arbitrary $v \in B_\epsilon(q) \cap T_p^\perp \cap B_\tau(p)$ where $q \in \bar{x}$ and $q \notin B_\epsilon(p)$. For such v the picture of fig. 1 can be drawn. Following lemma 4.1, we see that the distance of v to p is at most $\frac{\epsilon^2}{\tau}$. Now by the fact that $\bar{x}$ is $\frac{\epsilon}{2}$-dense, we have that there is some point $x \in \bar{x}$ which is within $\frac{\epsilon}{2}$ of p. The worst case picture of this is shown in fig 2. From lemma 4.2, we see that $v \in B_\epsilon(x)$ for this x. The proposition is proved.

□

These two propositions taken together show that $\mathcal{M}$ is a deformation retract of U. We see that $\mathcal{M} \subset U$. Further let $F(x,t) : U \times [0,1] \to U$ be given by $F(x,t) = tx + (1-t)\pi(x)$. Then F is continuous, $F(x,0) = \pi$ and $F(x,1)$ is the identity map.

Lemma 4.1 *Consider any $q \notin B_\epsilon(p)$. Let $v \in B_\epsilon(q) \cap T_p^\perp \cap B_\tau(p)$. Then the Euclidean distance from v to p is less than $\frac{\epsilon^2}{\tau}$.*

PROOF:It suffices to consider q on the curve as shown in fig. 1. Following the symbols on the figure, we have

$$A = b\sin(\theta) + \sqrt{\epsilon^2 - b^2\cos^2(\theta)}$$

where $b = 2\tau\sin(\theta)$. Therefore, we have

$$A = 2\tau\sin^2(\theta) + \sqrt{\epsilon^2 - 4\tau^2\sin^2(\theta)\cos^2(\theta)}$$

From this we see that

$$\frac{dA}{d\theta} = 2\tau\sin(2\theta) - \frac{4\tau^2\sin(2\theta)\cos(2\theta)}{2\sqrt{\epsilon^2 - \tau^2\sin^2(2\theta)}} = 2\tau\sin(2\theta)(1 - \frac{\tau\cos(2\theta)}{\sqrt{\epsilon^2 - \tau^2\sin^2(2\theta)}})$$

It is easy to check that if $\epsilon < \tau$, $\frac{dA}{d\theta} < 0$, i.e., A is monotonically decreasing with θ. Therefore the worst case situation is when $b = 2\tau\sin(\theta) = \epsilon$. For this value of θ, we see that $A = \frac{\epsilon^2}{\tau}$.

□

The following lemma ensures that there is an $x \in \bar{x} \cap B_\epsilon(p)$ such that $v \in B_\epsilon(x) \cap T_p^\perp$.

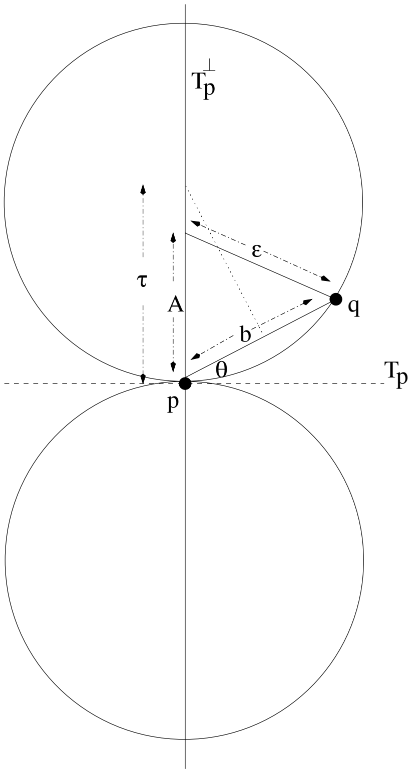

Figure 1: A picture showing the worst case.

Lemma 4.2 *Let $\bar{x}$ be $\epsilon/2$-dense in $\mathcal{M}$. For any $p \in \mathcal{M}$, let $v \in \pi^{-1}(p)$. Then for $0 < \epsilon < \sqrt{3/5}\tau$, we have that $v \in B_\epsilon(x) \cap T_p^\perp$ for some $x \in B_\epsilon(p) \cap \bar{x}$.*

PROOF:By the $\epsilon/2$ dense property, we know that there is an $x \in \bar{x}$ such that $x \in B_{\epsilon/2}(p)$. Consider the picture in fig. 2. This represents the most unfavorable position that such an x might have for the current context. By the same argument of lemma 4.1 we see that

$$A = \sqrt{\epsilon^2 - b^2\cos^2(\theta)} - b\sin(\theta)$$

where $b = 2\tau\sin(\theta) = \frac{\epsilon}{2}$. Putting this value in, we have

$$A = \sqrt{\epsilon^2 - \frac{\epsilon^2}{4}(1 - \frac{\epsilon^2}{16\tau^2})} - 2\tau\frac{\epsilon^2}{16\tau^2}$$

Simplifying, we see that $A > \frac{\epsilon^2}{\tau}$ if

$$\sqrt{\epsilon^2 - \frac{\epsilon^2}{4}(1 - \frac{\epsilon^2}{16\tau^2})} > \frac{9}{8}\frac{\epsilon^2}{\tau}$$

Squaring both sides, we have

$$\frac{3}{4}\epsilon^2 + \frac{\epsilon^4}{64\tau^2} > \frac{81\epsilon^4}{64\tau^2}$$

This simplifies to

$$\frac{\epsilon^2}{\tau^2} < \frac{3}{5}$$

Therefore, as long as $\epsilon < \sqrt{\frac{3}{5}}\tau$, we will have that $v \in B_\epsilon(x)$ for a suitable x. □

5 Probability Bounds

Following our assumption, that the points x_i are drawn at random, we now provide a bound on how many examples need to be drawn so that the empirically constructed complex has the same homology as the manifold. We begin with a basic probability lemma.

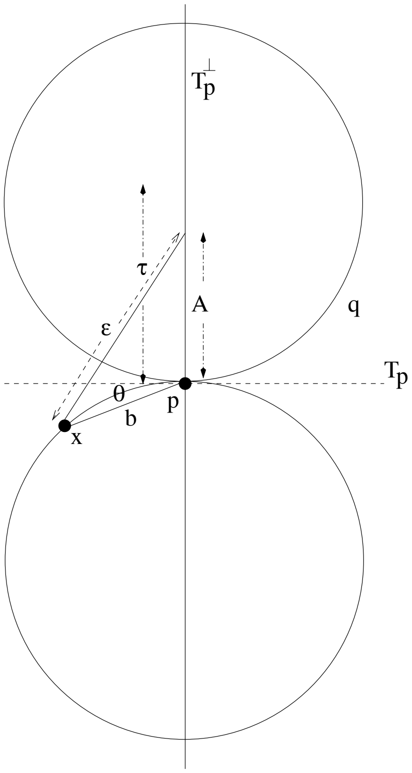

Figure 2: A picture showing the worst case.

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

$$n \geq \frac{1}{\alpha}\left(\log l + \log(\frac{1}{\delta})\right)$$

we are guaranteed that with probability $> 1 - \delta$, the following is true

$$\forall i, \bar{x} \cap A_i \neq \phi$$

PROOF:This follows from a simple application of the union bound. Let E_i be the event that $\bar{x} \cap A_i$ is empty. The probability with which this happens is given by

$$\mathbb{P} E_i = (1 - \mu(A_i))^n \leq (1 - \alpha)^n.$$

Therefore, by the union bound, we have

$$\mathbb{P} \cup_i E_i \leq \sum_{i=1}^{l} \mathbb{P} E_i \leq l(1 - \alpha)^n$$

It remains to show that for $n \geq \frac{1}{\alpha}\left(\log l + \log(\frac{1}{\delta})\right)$, we have

$$l(1 - \alpha)^n \leq \delta$$

To see this, simply note that $f(x) = xe^x - e^x + 1 \geq 0$ for all $x \geq 0$. This is seen by noting that $f(0) = 0$ and $f'(x) = xe^x \geq 0$ for all $x \geq 0$. Putting $x = \alpha$ in the above function, we have

$$(1 - \alpha) \leq e^{-\alpha}$$

and therefore it is easily seen that

$$l(1 - \alpha)^n \leq le^{-n\alpha} \leq \delta$$

for the appropriate choice of n.

□

Applying this to our setting, we consider a cover of the manifold $\mathcal{M}$ by balls of radius $\frac{\epsilon}{4}$. Let $\{y_i; 1 \leq i \leq l\}$ be the centers of such balls that constitute a minimal cover. Therefore, we can choose $A_i = B_{\frac{\epsilon}{4}}(y_i) \cap \mathcal{M}$. Applying the above lemma, we immediately have an estimate on the number of examples we need to collect. This is given by

$$\frac{1}{\alpha}\left(\log l + \log(\frac{1}{\delta})\right)$$

where

$$\alpha = \min_i \frac{vol(A_i)}{vol(\mathcal{M})}$$

and l is the $\frac{\epsilon}{4}$ covering number. These may be expressed entirely in terms of natural invariants of the manifold and we derive these quantities below. First, we note that the covering number may be bounded in terms of the packing number, i.e., the maximum number of sets of the form $N_i = B_r \cap \mathcal{M}$ (at scale r) that may be packed into $\mathcal{M}$ without overlap. In particular, if $C(\epsilon)$ is the ϵ-covering number of $\mathcal{M}$ and $P(\epsilon)$ is the ϵ-packing number, then the following simple lemma is true.

Lemma 5.2

$$P(2\epsilon) \leq C(2\epsilon) \leq P(\epsilon)$$

PROOF:The fact that $P(2\epsilon) \leq C(2\epsilon)$ follows from the definition. To see that $C(2\epsilon) \leq P(\epsilon)$, begin by letting $B_\epsilon(x_1), \ldots, B_\epsilon(x_N)$ be a a realization of an optimal ϵ-packing so that $N = P(\epsilon)$. We claim that $B_{2\epsilon}(x_1), \ldots, B_{2\epsilon}(x_N)$ form a 2ϵ-cover. If not, there exists an $x \in \mathcal{M}$ such that $B_\epsilon(x) \cap B_\epsilon(x_i)$ is empty for all i. In that case, one can add $B_\epsilon(x)$ to the collection to increase the packing number by 1 leading to a contradiction. Since $B_{2\epsilon}(x_1), \ldots, B_{2\epsilon}(x_N)$ is a valid 2ϵ-cover, we have $C(2\epsilon) \leq N = P(\epsilon)$.

□

Since l is the $\epsilon/4$ covering number, we see that $l \leq P(\epsilon/8)$ from lemma 5.2. Now we need to bound the packing number. To do so, we need the following result.

Lemma 5.3 *Let $p \in \mathcal{M}$. Now consider $A = \mathcal{M} \cap B_\epsilon(p)$. Then $vol(A) \geq (\cos(\theta))^k vol(B_\epsilon^k(p))$ where $B_\epsilon^k(p)$ is the k-dimensional ball centered at p, $\theta = \arcsin(\frac{\epsilon}{2\tau})$. All volumes are k-dimensional volumes.*

PROOF:Consider the tangent space at p given by T_p and let f be the projection to T_p. Let $B_r^k(p)$ be the k-dimensional ball of radius r centered at p lying in T_p. Let $f_A = \{f(q) \,|q \in A\}$ be the image of A under f. We will show that $B_r^k(p) \subset f_A$. Since f is a projection we have

$$vol(A) \geq vol(f_A) \geq vol(B_r^k(p)) = (\cos(\theta))^k vol(B_\epsilon^k(p))$$

To see that $B_r^k(p) \subset f_A$, notice that f is an open map whose derivative is nonsingular for all $q \in A$ (by Lemma 5.4). Therefore f is locally invertible and there exists a ball $B_s^k(p)$ of radius s such that $f^{-1}(B_s^k(p)) \subset A$. One can keep increasing s until it happens for the first time (say at $s = s'$) that $f^{-1}(B_s^k(p)) \not\subset A$. At this stage, there exists a point q in the closure of A such either (i) f is singular at q or (ii) $q \notin A$. By Lemma 5.4, we see that (i) is impossible. Therefore, $q \notin A$ but q is in the closure of A implying that $||q - p|| = \epsilon$. We see that $s' = \epsilon \cos(\phi)$ where ϕ is the angle between the line $\bar{qp}$ (the line joining q to p) and the line $\bar{f(q)p}$ (the line joining $f(q)$ to p). By the curvature bound implied by τ, we see that $|\phi| \leq |\theta|$ and therefore $s' = \epsilon \cos(\phi) \geq \epsilon \cos(\theta) = r$. □

Lemma 5.4 *Let $p \in \mathcal{M}$, let $A = \mathcal{M} \cap B_\epsilon(p)$, and let f be the projection to the tangent space at p (T_p). Then for all $\epsilon < \frac{\tau}{2}$, the derivative df is nonsingular at all points $q \in A$.*

PROOF:
Suppose df was singular for some $q \in A$. That means that the tangent space at q (T_q) is oriented so that the vector with origin q and end point $f(q)$ lies in T_q, in other words, T_q and T_p are at right angles to each other. Since $q \in B_\epsilon(p)$, we have that $d = ||q - p|| < \frac{\tau}{2}$. Putting propositions 6.2 and 6.3 together, we get that

$$\cos(\phi) \geq \sqrt{1 - \frac{2d}{\tau}} > 0$$

where ϕ is the angle between T_p and T_q. From this we see that $\phi < \frac{\pi}{2}$ leading to a contradiction.

□

Using lemma 5.3, we see that a simple bound on the packing number is obtained. We obtain immediately that

$$P(\epsilon) \leq \frac{vol(\mathcal{M})}{(\cos(\theta))^k vol(B_\epsilon^k(p))}$$

Therefore, we have

$$l \leq P(\frac{\epsilon}{8}) \leq \frac{vol(\mathcal{M})}{(\cos(\theta))^k vol(B^k_{\frac{\epsilon}{8}}(p))}$$

where $\theta = \arcsin(\frac{\epsilon}{16\tau})$. Similarly, we have that

$$\frac{1}{\alpha} \leq \frac{vol(\mathcal{M})}{(\cos(\theta))^k vol(B^k_{\epsilon}(p))}$$

6 Curvature and the Condition Number $\frac{1}{\tau}$

In this section[1], we examine the consequences of the condition number $\frac{1}{\tau}$ for the submanifold $\mathcal{M}$. As we have mentioned before, τ controls the curvature of the manifold at every point. This fact has been exploited in our earlier proofs. For submanifolds, one may formally study curvature through the second fundamental form (see e.g., [8]). Here we show formally that the norm of the second fundamental form is bounded by $\frac{1}{\tau}$. Thus a large τ corresponds to a well conditioned submanifold that has low curvature.

Proposition 6.1 states the bound on the norm of the second fundamental form. Proposition 6.2 states a bound on the maximum angle between tangent spaces at different points in $\mathcal{M}$. Proposition 6.3 states a bound on the maximum difference between the geodesic distance and the ambient distance for neighboring points in $\mathcal{M}$.

Let us begin by recalling the second fundamental form. Fix a point $p \in \mathcal{M}$. Following standard accounts (see, e.g. [8]), there exists a symmetric bilinear form $B : T_p \times T_p \to T_p^{\perp}$ that maps any two vectors in the tangent space ($u, v \in T_p$) into a vector $B(u,v)$ in the normal space. Thus for any normal vector (unit norm) $\eta \in T_p^{\perp}$, one can define the following

$$B_\eta(u,v) =< \eta, B(u,v) >=< u, L_\eta v >$$

where the inner product $< \cdot, \cdot >$ is the usual inner product in the tangent space of the ambient manifold (in our case $\mathbb{R}^k$). Since $B_\eta : T_p \times T_p \to \mathbb{R}$ is symmetric and bilinear, we see that $L_\eta : T_p \to T_p$ is a linear self-adjoint

[1] Thanks to Nat Smale for discussions leading to the writing of this section.

operator. The norm of the second fundamental form in direction η is now given by

$$\lambda_\eta = \sup_{u \in T_p} \frac{< u, L_\eta u >}{< u, u >}$$

It is seen that λ_η is the largest eigenvalue of L_η. Given this, we can prove the following proposition that characterizes the relation between the curvature through the second fundamental form and the condition number of the submanifold.

Proposition 6.1 *If $\mathcal{M}$ is a submanifold of $\mathbb{R}^k$ with condition number $\frac{1}{\tau}$, then the norm of the second fundamental form is bounded by $\frac{1}{\tau}$ in all directions. In other words, for all points $p \in \mathcal{M}$ and for all (unit norm) $\eta \in T_p^\perp$, we have*

$$\lambda_\eta = \sup_{u \in T_p} \frac{< u, L_\eta u >}{< u, u >} \leq \frac{1}{\tau}$$

PROOF:We prove by contradiction. Suppose the proposition is false. Then there exists a point $p \in \mathcal{M}$, a tangent vector (unit norm) $u \in T_p$ and a normal vector (unit norm) η such that

$$< \eta, B(u, u) > \quad > \quad \frac{1}{\tau}$$

Consider a curve $c(t) \in \mathcal{M}$ parametrized by arc length such that $c(0) = p$ and $\dot{c}(0) = \frac{dc}{dt}(0) = u$. For convenience, we will place the origin at p so that $c(0) = 0 = p$. With this (ambient) coordinate system, consider the point given by $\tau\eta$, i.e., the point a distance τ from p in the direction η. By our hypothesis on the condition number of the submanifold, we see that $p \in \mathcal{M}$ is the closest point on the manifold to the center of the τ-ball given by $\tau\eta$.

$$\text{for all } t, \quad ||c(t) - \tau\eta||^2 \geq \tau^2$$

from which we get

$$\forall t, \quad < c(t), c(t) > -2\tau < c(t), \eta > \quad \geq \quad 0$$

Consider the function $g(t) =< c(t), c(t) > -2\tau < c(t), \eta >$. Since $c(0) = 0$, we see that $g(0) = 0$. Further, we have $g'(t) = 2 < c(t), \dot{c}(t) > -2\tau <$

$\dot{c}(t), \eta >$. Since $c(0) = 0$ and $< \dot{c}(0), \eta >= 0$, we see that $g'(0) = 0$. Finally, $g''(t) = 2 < \dot{c}(t), \dot{c}(t) > +2 < c(t), \ddot{c}(t) > -2\tau < \ddot{c}(t), \eta >$. Since c is parameterized by arc length, we have $< \dot{c}(t), \dot{c}(t) > = 1$ and $g''(0) = 2 - 2\tau < \dot{c}(0), \eta >$.

Noting that the tangent vector field $\frac{dc}{dt}$ is parallel (see proof of Proposition 6.2), we see that $B(\frac{dc}{dt}, \frac{dc}{dt}) = \ddot{c}(t)$. Therefore, by assumption, we have that

$$< \eta, B(u,u) >=< \eta, B(\frac{dc}{dt}, \frac{dc}{dt}) >=< \eta, \ddot{c}(0) > \quad > \quad \frac{1}{\tau}$$

Therefore, $g''(0) < 2 - 2\tau(\frac{1}{\tau}) = 0$. By continuity, there exists a t^* such that $g(t^*) < 0$. But this leads to a contradiction since $g(t) \geq 0$ for all t.

□

Since the norm of the second fundamental form is bounded, we see that the manifold cannot curve too much locally. As a result, the angle between tangent spaces at nearby points cannot be too large. Let p and q be two points in the submanifold $\mathcal{M}$ with associated tangent spaces T_p and T_q. Since T_p and T_q are affine subspaces of $\mathbb{R}^k$, one can compare them in the ambient space in a standard way.

Formally, one may transport the tangent spaces to the origin (according to the standard connection defined in the ambient space $\mathbb{R}^k$) and then compare vectors in each of these tangent spaces with each other. Thus for any (unit norm) vectors $u \in T_p$ and $v \in T_q$, we may define the angle θ between them by

$$\cos(\theta) = | < u', v' > |$$

where $< \cdot, \cdot >$ is the usual inner product in $\mathbb{R}^k$, u', v' are the vectors obtained by parallel transport (in $\mathbb{R}^k$) of u and v respectively to the origin. Hereafter, we will always take this construction as standard. We will drop the prime notation and use $< u, v >$ to denote $< u', v' >$ in what follows.

We can now state the following proposition.

Proposition 6.2 *Let $\mathcal{M}$ be a submanifold of $\mathbb{R}^k$ with condition number $\frac{1}{\tau}$. Let $p, q \in \mathcal{M}$ be two points with geodesic distance given by $d_{\mathcal{M}}(p, q)$. Let ϕ be the the angle between the tangent spaces T_p and T_q defined by $\cos(\phi) = \min_{u \in T_p} \max_{v \in T_q} | < u, v > |$. Then $\cos(\phi)$ is greater than $1 - \frac{1}{\tau} d_{\mathcal{M}}(p, q)$.*

PROOF:Consider two points $p, q \in \mathcal{M}$ connected by a geodesic curve $c(t) \in \mathcal{M}$. Let $c(t)$ be parametrized (proportional to arc length) so that $c(0) = p$, and $c(1) = q$.

Now let $v_p \in T_p$ be a tangent vector (unit norm) and let $v(t)$ be the parallel transport of this vector along the curve $c(t)$. Thus we have $v(0) = v_p$, $v(1) = v_q \in T_q$. Clearly $< v(t), v(t) >= 1$ for all t since v is parallel.
Notice that

$$< v(0), v(1) >=< v(0), v(0) + w >= 1+ < v(0), w > \tag{1}$$

where

$$w = \int_0^1 (\frac{dv}{dt})dt \tag{2}$$

Combining 1 and 2, we see

$$\cos(\theta) = | < v(0), v(1) > | \geq 1 - | < v(0), w > | \geq 1 - ||w|| \tag{3}$$

where θ is the angle between the vectors $v(0)$ and $v(1)$. Since $v_p = v(0)$ was arbitrary, it is easy to check that $\cos(\phi) \geq \cos(\theta)$.
Now

$$\frac{dv}{dt} = \bar{\nabla}_{\frac{dc}{dt}} v(t)$$

where $\bar{\nabla}$ denotes the connection in Euclidean space. At the same time

$$\nabla_{\frac{dc}{dt}} v(t) = (\bar{\nabla}_{\frac{dc}{dt}} v(t))^T$$

where for any $r \in \mathcal{M}$ and $v \in \bar{T}_r$ (here $\bar{T}_r$ is the tangent space of $\mathbb{R}^k$ at r) we denote by $(v)^T$ the projection of v onto T_r (here T_r is the tangent space to $\mathcal{M}$ at r viewed as an affine space with origin r). But since $v(t)$ is parallel, we have that $\nabla_{\frac{dc}{dt}} v(t) = 0$. Therefore, $\bar{\nabla}_{\frac{dc}{dt}} v(t)$ is entirely in the space normal to $T_{c(t)}$. But the component of $\bar{\nabla}_{\frac{dc}{dt}} v(t)$ in the normal direction is precisely given by the second fundamental form. Hence, we have that

$$\frac{dv}{dt} = B(\frac{dc}{dt}, v(t))$$

where B is a symmetric, bilinear form (the second fundamental form). Letting η be a unit norm vector in the direction $\frac{dv}{dt}$, i.e., $\eta = (1/||\frac{dv}{dt}||)\frac{dv}{dt}$, we see that

$$||\frac{dv}{dt}|| =< \eta, \frac{dv}{dt} >=< \eta, B(\frac{dc}{dt}, v(t)) =< \frac{dc}{dt}, L_n v(t) >$$

where L_n is a self adjoint linear operator. By Proposition 6.1, the norm of L_η is bounded by $\frac{1}{\tau}$. Therefore, we have

$$||\frac{dv}{dt}|| \le ||\frac{dc}{dt}||||L_n v|| \le ||\frac{dc}{dt}||||L_\eta||$$

and

$$||w|| = ||\int_0^1 \frac{dv}{dt}|| \le \int_0^1 ||\frac{dv}{dt}|| \le ||L_n|| \int_0^1 ||\frac{dc}{dt}||dt \le \frac{1}{\tau} d_{\mathcal{M}}(p,q) \qquad (4)$$

Combining eq. 3 and eq. 4, we get

$$\cos(\phi) \ge 1 - \frac{1}{\tau} d_{\mathcal{M}}(p,q)$$

□

We next show a relationship between the geodesic distance $d_{\mathcal{M}}(p,q)$ and the ambient distance $||p-q||_{\rm I\!R^k}$ for any two points p and q on the submanifold $\mathcal{M}$.

Proposition 6.3 *Let $\mathcal{M}$ be a submanifold of ${\rm I\!R}^k$ with condition number $\frac{1}{\tau}$. Let p and q be two points in $\mathcal{M}$ such that $||p-q||_{\rm I\!R^k} = d$. Then for all $d \le \frac{\tau}{2}$, the geodesic distance $d_{\mathcal{M}}(p,q)$ is bounded by*

$$d_{\mathcal{M}}(p,q) \le \tau - \tau\sqrt{1 - \frac{2d}{\tau}}$$

PROOF:Consider two points $p, q \in \mathcal{M}$ and let $c(t)$ be a geodesic curve joining them such that $c(0) = p$ and $c(s) = q$. Let c be parametrized by arc length so that $||\dot{c}(t)|| = 1$ for all t and $d_{\mathcal{M}}(p,q) = s$.
Noting that the tangent vector field $\dot{c}$ along the curve is parallel, we have $\ddot{c} = B(\dot{c}, \dot{c})$ and from proposition 6.1, we see that for all t

$$||\ddot{c}|| = ||B(\dot{c}, \dot{c})|| \le \frac{1}{\tau}$$

The chord length between p and q is given by $||c(s) - c(0)||$ and we now relate this to the geodesic distance $d_{\mathcal{M}}(p,q)$. Observe that

$$c(s) - c(0) = \int_0^s \dot{c}(t)dt$$

Now

$$\dot{c}(t) = \dot{c}(0) + \int_0^t \ddot{c}(r)dr$$

Thus $\dot{c}(t) = \dot{c}(0) + u(t)$ where $u(t) = \int_0^t \ddot{c}(r)dr$. We see that

$$||u(t)|| \leq \int_0^t ||\ddot{c}(r)dr|| \leq \frac{t}{\tau}$$

Therefore,

$$||c(s)-c(0)|| = || \int_0^s \dot{c}(0)dt + \int_0^s u(t)dt|| \geq s||\dot{c}(0)|| - \int_0^s ||u(t)||dt \geq s - \int_0^s \frac{t}{\tau}dt$$

Therefore we get

$$||c(s) - c(0)|| = d \geq s - \frac{s^2}{2\tau} \tag{5}$$

where d is the ambient distance between the points p and q while s is the geodesic distance between these same points. The inequality in eq. 5 is satisfied only if $s \leq \tau - \tau\sqrt{1 - \frac{2d}{\tau}}$ or $s \geq \tau + \tau\sqrt{1 - \frac{2d}{\tau}}$. Since $s = 0$ when $d = 0$, we know that the second inequality does not apply. Therefore, from the first inequality, we have

$$s \leq \tau - \tau\sqrt{1 - \frac{2d}{\tau}}$$

□

7 Handling Noisy Data

In this section we show that if our data is noisy in the sense that it is drawn from a probability distribution that is concentrated around (rather than on) the manifold, the homology of the manifold can still be computed from noisy data.

7.1 The Model of Noise

Consider a probability measure μ concentrated around the manifold. We assume that μ satisfies the following two regularity conditions.

1. The support of μ (suppμ) is contained in the tubular neighborhood of radius r around $\mathcal{M}$. Thus supp$\mu \subset \text{Tub}_r(\mathcal{M})$.
2. For every $0 < s < r$, we have that

   $$\inf_{p \in \mathcal{M}} \mu(B_s(p)) > k_s$$

   where k_s is a constant depending on s and independent of p.

In what follows, we assume the data is drawn in i.i.d. fashion according to a P that satisfies the above properties.

7.2 Main Topological Lemma: Sufficient Conditions

We will proceed by constructing ϵ-balls centered on our data points. If these data are s-dense on the manifold, then the homology of the union of these balls will equal that of the manifold $\mathcal{M}$ *even if* the data is drawn from a noisy distribution. In order to see that this might be the case at all, we provide a simple argument. This argument works with non-optimal choices of ϵ and s and later sections will enter into the considerations of choosing better values for these parameters and therefore provide more natural complexity estimates.

Let $\bar{x} = \{x_1, \ldots, x_n\}$ be a set of n points in the tubular neighborhood of radius r around $\mathcal{M}$. Let U be given by

$$U = \cup_{x \in \bar{x}} B_\epsilon(x)$$

Proposition 7.1 *If $\bar{x}$ is r-dense in $\mathcal{M}$ then $\mathcal{M}$ is a deformation retract of U for all $r < (\sqrt{9} - \sqrt{8})\tau$ and $\epsilon \in \left(\frac{(r+\tau)-\sqrt{r^2+\tau^2-6\tau}}{2}, \frac{(r+\tau)+\sqrt{r^2+\tau^2-6\tau}}{2}\right)$.*

PROOF:We show that for each $p \in \mathcal{M}$, it is the case that $\pi^{-1}(p)$ contracts to p. Consider a $v \in \pi^{-1}(p)$. Consider the line segment, $\bar{vp}$, joining v to p. We claim that this line segment is entirely contained in $\pi^{-1}(p)$. Clearly, if $v \in B_\epsilon(x)$ for some $x \in \bar{x} \cap B_\epsilon(p)$, this is immediate by the convexity of balls in Euclidean space. So we only need to consider the situation where $v \in B_\epsilon(x)$ for some $x \notin \bar{x} \cap B_\epsilon(p)$. So let $v \in B_\epsilon(q) \cap T_p^\perp$. Let

$$u = \arg \min_{x \in \bar{vp} \cap \bar{B_\epsilon(q)}} ||x - p||$$

As long as $u \in B_\epsilon(x)$ for some $x \in \bar{x} \cap B_\epsilon(p)$, we see that the line segment $\bar{up}$ is contained in $\pi^{-1}(p)$ and therefore v contracts to p.
Since we choose $r < \epsilon$, we are guaranteed that there is an $x \in \bar{x} \cap B_r(p) \subset B_\epsilon(p)$. The worst case picture is shown in fig. 3. Following the symbols of the picture, as long as

$$\tau - A < \epsilon - r,$$

we have that v contracts to p. Thus we need

$$(\tau - (\epsilon - r))^2 < A^2 = (\tau - r)^2 - \epsilon^2 \tag{6}$$

Expanding the squares, this reduces to

$$\epsilon^2 - \epsilon(\tau + r) + 2\tau r < 0$$

This is a quadratic in ϵ and is satisfied for

$$\epsilon \in \left(\frac{(r+\tau) - \sqrt{r^2 + \tau^2 - 6\tau r}}{2}, \frac{(r+\tau) + \sqrt{r^2 + \tau^2 - 6\tau r}}{2} \right) \tag{7}$$

provided

$$r^2 - 6\tau r + \tau^2 > 0$$

This, in turn, is a quadratic in r and it is easy to check that it is satisfied as long as

$$r < (3 - 2\sqrt{2})\tau = (\sqrt{9} - \sqrt{8})\tau \tag{8}$$

Thus we see that for r, ϵ satisfying equations 7 and 8, we have that v contracts to p. □

We now need to compute the probability of drawing a random $\bar{x}$ that is guaranteed to be r-dense. The following proposition is true.

Proposition 7.2 *Let $N_{r/2}$ be the $r/2$-covering number of the manifold. Let $p_1, \ldots, p_{N_{r/2}} \in \mathcal{M}$ be points on the manifold such that $B_{r/2}(p_i)$ realize an $r/2$-cover of the manifold. Let $\bar{x}$ be generated by i.i.d. draws according to a probability measure μ that satisfies the regularity properties described earlier. Then if if $|\bar{x}| > \frac{1}{k_{r/2}} \left(\log(N_{r/2}) + \log(\frac{1}{\delta}) \right)$, with probability greater than $1 - \delta$, $\bar{x}$ will be r-dense in $\mathcal{M}$.*

PROOF:Take $A_i = B_{r/2}(p_i)$ and apply Lemma 5.1. By the conclusion of that lemma, we have that with high probability each of the A_i's is occupied by atleast one $x \in \bar{x}$. Therefore it follows that for any $p \in \mathcal{M}$, there is atleast

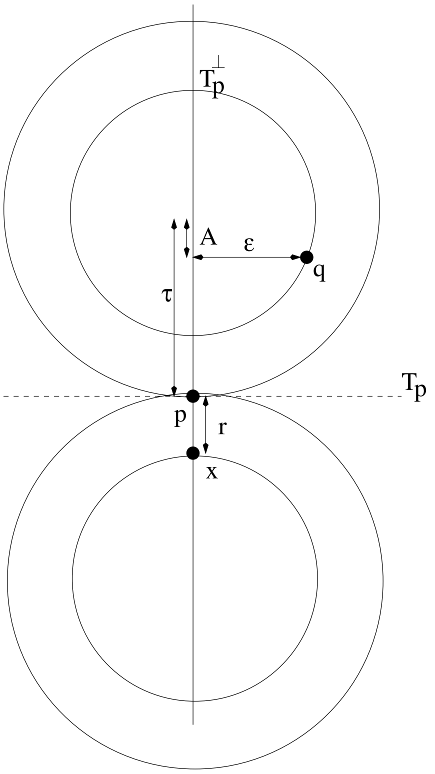

Figure 3: A picture showing the worst case.

one $x \in \bar{x}$ such that $||p - x|| < r$. Thus with high probability $\bar{x}$ is r-dense on the manifold.

□

Putting these together, our main conclusion is

Theorem 7.1 *Let $N_{r/2}$ be the $r/2$-covering number of the submanifold $\mathcal{M}$ of $\mathbb{R}^k$. Let $\bar{x}$ be generated by i.i.d. draws according to a probability measure μ that satisfies the regularity properties described earlier. Let $U = \cup_{x \in \bar{x}} B_\epsilon(x)$. Then if if $|\bar{x}| > \frac{1}{k}\left(\log(N_{r/2}) + \log(\frac{1}{\delta})\right)$, with probability greater than $1 - \delta$, $\mathcal{M}$ is a deformation retract of U as long as (1) $r < (\sqrt{9} - \sqrt{8})\tau$ and (2) $\epsilon \in \left(\frac{(r+\tau) - \sqrt{r^2+\tau^2-6\tau r}}{2}, \frac{(r+\tau) + \sqrt{r^2+\tau^2-6\tau r}}{2}\right)$*

7.3 Main Topological Lemma – General Considerations

In general, we may demand points that are s-dense. Putting ϵ-balls around these points we construct U in the usual way. The condition number τ and the noise bound r are additional parameters that are outside our control and determined externally. We now ask what is the feasible space (s, ϵ, r, τ) that will guarantee that U is homotopy equivalent to $\mathcal{M}$?

Following our usual logic, we see that the worst case situation is given by fig. 4. An arbitrary $v \in B_\epsilon(q) \cap T_p^\perp \cap B_\tau(p)$ will contract to p if

$$B_\epsilon(q) \cap B_\epsilon(x) \cap \bar{vp} \neq \phi$$

This is the same as requiring

$$(\tau - v)^2 < (\tau - r)^2 - \epsilon^2 \tag{9}$$

Additionally, we have the following equations that need to be satisfied (following fig. 4).

$$(\tau - r)^2 - (\tau - \beta)^2 = s^2 - \beta^2 \tag{10}$$

$$s^2 - \beta^2 + (\beta + v)^2 = \epsilon^2 \tag{11}$$

If one eliminates v and β from the above equations, one will get a single inequality relating s, ϵ, τ, r that describes for each τ, r the feasible set of possible choices of s, ϵ that are sufficient to guarantee homotopy equivalence. Let us see how our earlier theorems follow from particular choices of this general set of equations.

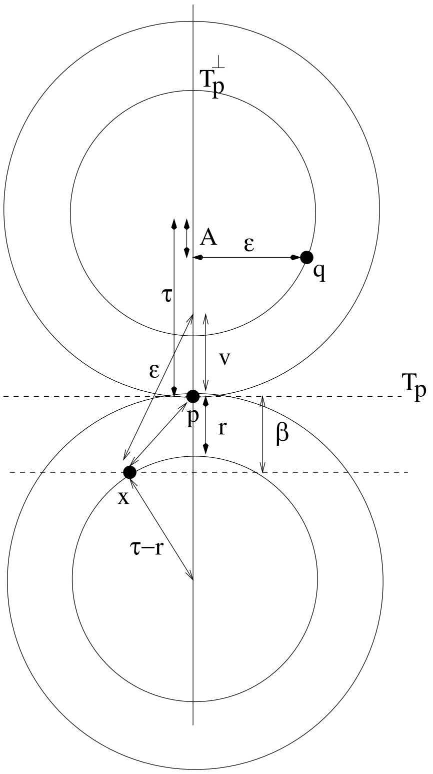

Figure 4: A picture showing the worst case.

7.3.1 The Case when $s = r$

We have already examined the case when the points $\bar{x}$ are chosen to be r-dense in $\mathcal{M}$. Putting in $s = r$ in equations 9, 10, and 11, we see the following:
From eq. 10, we have (for $s = r$)

$$(\tau - r)^2 - (\tau - \beta)^2 = r^2 - \beta^2$$

This simplifies to give $\beta = r$.
Putting $\beta = r$ and $s = r$ in eq. 11, we get

$$r^2 - r^2 + (r + v)^2 = \epsilon^2$$

giving us $v = \epsilon - r$.
Finally, putting $v = \epsilon - r$ in inequality 9, we get

$$(\tau - (\epsilon - r))^2 < (\tau - r)^2 - \epsilon^2$$

which is the same as inequality 6 whose solution was examined in the previous section.

7.3.2 The Case when $r = 0$

We can recover our main theorem for the noise-free case by considering the case $r = 0$. We proceed to do this now.
The fundamental inequality of 9 gives us (for $r = 0$)

$$(\tau - v)^2 < \tau^2 - \epsilon^2$$

This is the same as requiring

$$v^2 - 2\tau + \epsilon^2 < 0$$

Using standard analysis for quadratic functions, we see that the following condition is required:

$$v > \tau - \sqrt{\tau^2 - \epsilon^2} \tag{12}$$

We can eliminate v using equations 10 and 11. Thus, from eq. 10, we get $\beta = \frac{s^2}{2\tau}$ and substituting in eq. 11, we get a quadratic equation in v whose positive solution is given by $v = -\frac{s^2}{2\tau} + \sqrt{\frac{s^4}{4\tau^2} + (\epsilon^2 - s^2)}$. This gives rise to the following condition

$$-\frac{s^2}{2\tau} + \sqrt{\frac{s^4}{4\tau^2} + (\epsilon^2 - s^2)} > \tau - \sqrt{\tau^2 - \epsilon^2} \qquad (13)$$

Inequality 13 gives the feasible region for s and ϵ for the homotopy equivalence of U and $\mathcal{M}$. Let us consider the special case when $s = \frac{\epsilon}{2}$ — a choice we made in Section 3 without any attention to optimality. Putting in this value, after several simplifying steps, one obtains that

$$\epsilon^4 + 51\epsilon^2\tau^2 - 48\tau^4 < 0 \qquad (14)$$

This is satisfied for all $0 < \epsilon^2 < 0.9244\tau^2$ or

$$0 < \epsilon < 0.96\tau$$

Remark 1 Note that in our original proof of our main noise free theorem (Theorem 3.1), the deformation retract argument of Section 3 passes through the construction of $st(p)$ and shows contraction of $\pi^{-1}(p)$ by equating it with $st(p)$. This condition is stronger than we require. Here we see that the condition $B_\epsilon(q) \cap B_\epsilon(x) \cap \bar{vp} \neq \phi$ is sufficient. This latter condition is weaker and therefore gives us a slightly stronger version of Theorem 3.1 in the sense that it holds for a larger range of ϵ.

Remark 2 If we assume that τ, r are beyond our control, the sample complexity depends entirely upon s. Therefore if we wish to proceed by drawing the fewest number of examples, then it is necessary to maximize s subject to the condition of eq. 13.

Remark 3 The total complexity of finding the homology depends both upon s and ϵ in a more complicated way. The size of $\bar{x}$ depends entirely upon s and nothing else. However, the number of k-tuples to consider in the simplicial complex depends both upon the size of $\bar{x}$ as well as ϵ because ϵ determines how many balls will have non-empty intersections. We leave this more nuanced complexity analysis for future consideration.

References

[1] A. Zomorodian and G. Carlsson. Computing persistent homology. 20th ACM Symposium on Computational Geometry, Brooklyn, NY, June 9-11, 2004.

[2] J. B. Tenenbaum, V. De Silva, J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. Science 290 (5500): 22 December 2000.

[3] M. Belkin and P. Niyogi. Semisupervised Learning on Riemannian Manifolds. Machine Learning. Vol. 56. 2004.

[4] A. Bjorner. Topological Methods. in "Handbook of Combinatorics", (Graham, Grotschel, Lovasz (ed.)), North Holland, Amsterdam, 1995.

[5] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. Science 290: 2323-2326.2000.

[6] D. Donoho and C. Grimes. Hessian Eigenmaps: New Locally-Linear Embedding Techniques for High-Dimensional Data. Preprint. Stanford University, Department of Statistics. 2003.

[7] T. K. Dey, H. Edelsbrunner and S. Guha. Computational topology. Advances in Discrete and Computational Geometry, 109-143, eds.: B. Chazelle, J. E. Goodman and R. Pollack, Contemporary Mathematics 223, AMS, Providence, 1999.

[8] M. P. Do Carmo. Riemannian Geometry. Birkhauser. 1992.

[9] J. Munkres. Elements of Algebraic Topology. Perseus Publishing. 1984.

[10] J. Friedman. Computing Betti Numbers via the Combinatorial Laplacian. *Algorithmica*. 21. 1998.

[11] K. Fischer, B. Gaertner, M. Kutz. Fast Smallest-Enclosing-Ball Computation in High Dimensions. Proc. 11th Annual European Symposium on Algorithms (ESA), 2003.

[12] Website for Smallest Enclosing Ball Algorithm. http://www2.inf.ethz.ch/personal/gaertner/miniball.html

[13] T. Kaczynski, K. Mischaikow, M. Mrozek. Computational Homology. Springer Verlag, NY. Vol. 157. 2004.