Discrete Comput Geom 32:245–267 (2004)
DOI: 10.1007/s00454-004-1120-8

Discrete & Computational
Geometry
© 2004 Springer-Verlag New York, LLC

Shape Dimension and Intrinsic Metric from Samples of Manifolds*

Joachim Giesen and Uli Wagner

Institut für Theoretische Informatik, ETH Zürich,
CH-8092 Zürich, Switzerland
{giesen,uli}@inf.ethz.ch

Abstract. We introduce the *adaptive neighborhood graph* as a data structure for modeling a smooth manifold M embedded in some Euclidean space $\mathbb{R}^d$. We assume that M is known to us only through a finite sample $P \subset M$, as is often the case in applications. The adaptive neighborhood graph is a geometric graph on P. Its complexity is at most $\min\{2^{O(k)}n, n^2\}$, where $n = |P|$ and $k = \dim M$, as opposed to the $n^{\lceil d/2 \rceil}$ complexity of the Delaunay triangulation, which is often used to model manifolds. We prove that we can correctly infer the connected components and the dimension of M from the adaptive neighborhood graph provided a certain standard sampling condition is fulfilled. The running time of the dimension detection algorithm is $d2^{O(k^7 \log k)}$ for each connected component of M. If the dimension is considered constant, this is a constant-time operation, and the adaptive neighborhood graph is of linear size. Moreover, the exponential dependence of the constants is only on the intrinsic dimension k, not on the ambient dimension d. This is of particular interest if the co-dimension is high, i.e., if k is much smaller than d, as is the case in many applications. The adaptive neighborhood graph also allows us to approximate the geodesic distances between the points in P.

1. Introduction

Manifold learning is the problem of computing a model of a k-dimensional manifold M embedded in d-dimensional Euclidean space $\mathbb{R}^d$ only from a finite set P of sample points. Often k is very small compared with d.

The manifold learning problem was identified as one of the most important and challenging problems in computational topology during an NSF founded workshop on computational topology [7]. The importance of the problem can be stressed by its many applications, e.g., in speech recognition, weather forecasting, and economic prediction.

* The first author was partially supported by the Swiss National Science Foundation under the project “Non-linear manifold learning”.

The term manifold learning was introduced in 1995 in the context of speech understanding [9], [10]. However, the problem also appeared in research on neural networks [17] and in mathematical psychology [19]. In all these areas heuristics were developed to address the problem, but so far no theoretical correctness guarantees could be proven.

Special cases of the manifold learning problem in low dimensions received a lot of attention in recent years. Most prominent is the surface reconstruction problem where P is sampled from a two-dimensional manifold embedded in three-dimensional space. The surface reconstruction problem as described here is also special in the sense that the co-dimension of the problem, i.e., $d - k$, is one. Some of the proposed algorithms for curve and surface reconstruction [2]–[4], [11] come with the guarantee that the computed model is a manifold homeomorphic and geometrically close to the unknown manifold M provided some sampling condition is fulfilled. These algorithms are called provably correct. Almost all known provably correct algorithms for surface reconstruction use the three-dimensional Delaunay triangulation or the Voronoi diagram of the sample points as a basic data structure. Unfortunately the time to compute the Delaunay triangulation or the Voronoi diagram in dimension d is of the order $n^{\lceil d/2 \rceil}$ in the worst case, where $n = |P|$ is the number of sample points.

Point sets that occur "in practice" do not seem to exhibit this worst-case behavior, and recently it was shown [6], [13] that for points that are "nicely distributed" on a smooth surface in $\mathbb{R}^3$, the Delaunay triangulation has complexity much lower than n^2 (see [13] for an excellent survey of the different ways to specify "nice" and of related work). Observe, however, that exponential growth in the co-dimension is inevitable: As the simplest example, consider k mutually orthogonal unit circles centered at the origin in $\mathbb{R}^{2k}$. Even if we take n/k evenly spaced points on each of the circles, the complexity of the Delaunay triangulation will be $\Omega(n^k)$. Because of this exponential growth of the running time, the use of the Delaunay triangulation is prohibitive for high co-dimensions.

Even though the provably correct algorithms for curve and surface reconstruction cannot be adapted directly for use in higher dimensions, many of the ideas developed for their correctness proofs can. Especially interesting are the *sampling conditions* used in the proofs. Amenta and Bern [2] introduced the notion of an ε-sample, which can be defined in arbitrary dimensions and allows for non-uniform sampling. Later, the ε-sampling condition was specialized in [12] and [14] to the so-called (ε,δ)-*sampling condition*. We review these sampling conditions in Section 2, and then work with the (ε, δ)-sampling condition.

In this paper we introduce the *adaptive neighborhood graph*. The adaptive neighborhood graph is a geometric graph on the sample P and can be computed in time $O(n^2)$, without exponential dependence on d. We show that it is a provably good model for the manifold M in the following sense:

(1) If the sample P satisfies an (ε, δ)-sampling condition, then the adaptive neighborhood graph has the same connectivity as the manifold M, i.e., the points in P that are connected by a path in M are also connected by a path in the adaptive neighborhood graph and vice versa.

(2) Under the same sampling condition we can correctly infer the dimension k of M in time $d2^{O(k^7 \log k)}$ for each connected component. If the dimension is considered

a constant, then this is a constant-time operation, and, moreover, the exponential dependence of the constants is only on k and not on d. This improves a result of Dey et al. [12] who gave a dimension detection algorithm which is provably correct under the same sampling condition and is based on the d-dimensional Delaunay triangulation, i.e., requires time $\Theta(n^{\lceil d/2 \rceil})$ in the worst case. Dey et al. also provide an example that the sampling condition cannot be weakened essentially.

(3) The geodesic distance between two sample points on M can be approximated by the length of the shortest path that connects them in the adaptive neighborhood graph. That is, one can use Dijkstra's all pairs shortest path algorithm to approximate geodesic distances. This also improves an early result: Tenenbaum et al. [19] suggest two different graphs models: either they connect two points if they are at a distance of no more than some global threshold, or they use a fixed number of nearest neighbors around each point. The disadvantage of the first method is that it requires a much stricter sampling condition (globally uniform samples). The second graph, on the other hand, does not automatically adapt to the dimension of M. While for various input models (such as (ε, δ)-samples reviewed below, or uniform random sample points), one can show that there is a number m, depending on the model and the intrinsic dimension k, such that considering the m nearest neighbors works well for that model and that k, this number would have to be a parameter of the algorithm, to be interactively adjusted by the user for each particular input.

2. Sampling and the Adaptive Neighborhood Graph

Let $\mathcal{M} = \{M_1, \ldots, M_m\}$ be a collection of disjoint, smooth (C^2 would be enough), compact, and connected manifolds embedded in $\mathbb{R}^d$ and let $M = \bigcup_{i=1}^m M_i$ be the underlying topological space of $\mathcal{M}$. We do not assume that all the manifolds $M_i \in \mathcal{M}$ have the same dimension, but we assume that all the manifolds M_i have dimension larger than zero. Note that we assume our manifolds to have no boundary.

For $p \in M$, we denote by T_pM the tangent space of M at p, and by N_pM the space of normals. The tangent space T_pM is the set of tangent vectors at p of all smooth curves in M passing through p. If the dimension of M is k, then T_pM is a k-flat containing p. The space of normals N_pM is the orthogonal complement of T_pM. (We picture both spaces to be anchored at p.)

If we are to infer any useful information about M from a finite sample $P \subset M$, the sample has to fulfill some sampling condition. Following [2] we base the sampling condition on the medial axis of M.

Medial Axis and Local Feature Size. A d-dimensional ball B is called a *medial ball* of M if $\mathrm{int}(B) \cap M = \emptyset$ and $|\mathrm{bd}(B) \cap M| \geq 2$, i.e., B does not contain any point of M in its interior but at least two points of M in its boundary. The *medial axis* of M is the closure of the set of all medial ball centers.

The *local feature size* is the function $f\colon M \to \mathbb{R}$ that assigns to $x \in M$ its distance to the medial axis. The following observation was made in [2].

Lemma 1. *The feature size* $f\colon M \to \mathbb{R}$ *satisfies*:

(i) $f(x) \leq f(y) + \|x - y\|$.
(ii) *If* $\|x - y\| \leq \varepsilon f(x)$ *with* $\varepsilon < 1$, *then*

$$\|x - y\| \leq \frac{\varepsilon}{1 - \varepsilon} f(y).$$

In what follows, we assume that $f(x) < \infty$ for all $x \in M$. Observe that it follows from smoothness that $f(x) > 0$.

Most algorithms learning properties of M from a sampling P are based on the assumption that P is a uniform sample. This assumption is quite strict. It means that the sampling density is globally determined by the smallest feature exhibited by M. We use the feature size to define less restrictive sampling conditions.

Sampling Conditions. Let $\varepsilon > 0$. A finite sample $P \subset M$ is called an ε-sample if

$$\forall x \in M, \quad \exists p \in P, \qquad \|x - p\| \leq \varepsilon f(x).$$

An ε-sample P is called an (ε, δ)-sample or *tight* ε-sample if it satisfies the additional condition

$$\forall p, q \in P, \qquad \|p - q\| \geq \delta f(p)$$

for some δ, $0 < \delta < \varepsilon$.

The ε-sampling condition was introduced by Amenta et al. in the context of curve and surface reconstruction [2]. The tight ε-sampling condition was introduced by Dey et al. for dimension detection from samples [12] and by Funke and Ramos for fast surface reconstruction [14].

The fundamental data structure that we use later is the adaptive neighborhood graph.

Adaptive Neighborhood Graph. For a constant $c > 1$ we define the c-neighborhood $N_c(p)$ of a sample point $p \in P$ as follows:

$$N_c(p) = \left\{ q \in P - \{p\} \,:\, \|p - q\| \leq c \min_{q' \in P - \{p\}} \|p - q'\| \right\}.$$

The adaptive neighborhood graph $G_c(P)$ is the geometric graph with vertex set P where two vertices $p, q \in P$ are connected by a straight edge if either $q \in N_c(P)$ or $p \in N_c(q)$.

3. Connectivity

In this section we show that for a suitable choice of c and all sufficiently small $\varepsilon, \delta > 0$ such that the ratio δ/ε is at least some suitable constant, the adaptive neighborhood graph $G_c(P)$ of any (ε, δ)-sample P of M has essentially the same connectivity as M. That is, two points in P are connected by a path in M if and only if they lie in the same connected component of $G_c(P)$.

We start by proving an upper bound on the distance of a sample point from its nearest neighbor in the set of sample points. Moreover, we establish an upper bound on the length of restricted Delaunay edges. The *Delaunay triangulation of P restricted to M* is the dual complex of the *restricted Voronoi diagram of P*. That is, the convex hull of sample points $p_1, \ldots, p_k \in P$ belongs to the restricted Delaunay triangulation iff the intersection $M \cap \bigcap_{i=1}^{k} V_{p_i}$ is non-empty, where $V_p := \{x \in \mathbb{R}^d : \|x - p\| = \min_{q \in P} \|x - q\|\}$ is the Voronoi cell of a sample point p.

Lemma 2. *Let P be an ε-sample of M with $\varepsilon < \frac{1}{2}$. Then the following hold*:

(i) *The distance between $p \in P$ and its nearest neighbor in $P \backslash \{p\}$ is at most $(2\varepsilon/(1-\varepsilon)) f(p)$.*

(ii) *If $p, q \in P$ and if pq is an edge of the restricted Delaunay triangulation, then $\|p - q\| \leq (2\varepsilon/(1-\varepsilon)) \min\{f(p), f(q)\}$.*

Proof. Let V_p be the Voronoi cell of p. We first show that the boundary of V_p must have a non-empty intersection with the component M_i of M containing p.

Otherwise, M_i has to be completely contained in the interior of V_p and p is the only sample point on M_i. By compactness, there exists a point $x \in M_i$ whose distance to p is maximal. Observe that the tangent space $T_x M$ must be orthogonal to the segment xp, else a suitable small perturbation would produce a point even farther away from p. Thus, there exists a ball B of radius $f(x)$ that is tangent to x and whose center lies on the ray from x through p. The interior of B contains no point from M, in particular, the distance from p to x is at least the diameter of B, i.e., $\|x - p\| \geq 2f(x)$. Thus, if p is the only sample point on M_i, then S is not an ε-sample.

Thus, we can assume that V_p contains a point $y \in M_i$ in its boundary. From $\|y - p\| \leq \varepsilon f(y)$ and the property of the local feature size stated in Lemma 1(ii) we conclude that

$$\|y - p\| \leq \frac{\varepsilon}{1-\varepsilon} f(p).$$

Since y is contained in the boundary of V_p there must exist another point $q \in P \backslash \{p\}$ such that y is also contained in the boundary of the Voronoi cell V_q. We have

$$\|p - y\| = \|q - y\| \geq \frac{\|p - q\|}{2}$$

and thus

$$\|p - q\| \leq 2\|p - y\| \leq \frac{2\varepsilon}{1-\varepsilon} f(p).$$

Hence p has a neighbor in $P \backslash \{p\}$ within distance at most $(2\varepsilon/(1-\varepsilon)) f(p)$. This proves the first claim. Note that the edge pq is a restricted Delaunay edge. In general we have for any restricted Delaunay edge pq that the common intersection of M with the Voronoi regions V_p and V_q is not empty. Thus we can use the same calculations as above to prove also the second claim. □

Corollary 1. *Let P be an (ε, δ)-sample of M. If pq with $p, q \in P$ is an edge in the Delaunay triangulation of P restricted to M, then*

$$p \in N_c(q) \quad \text{and} \quad q \in N_c(p), \qquad \text{provided} \quad c \geq \frac{2\varepsilon}{(1-\varepsilon)\delta}.$$

Proof. Lemma 2(ii) together with the δ-condition satisfied by P gives

$$\begin{aligned} \|p - q\| &\leq \frac{2\varepsilon}{1-\varepsilon} f(p) \\ &\leq \frac{2\varepsilon}{(1-\varepsilon)\,\delta} \min_{p' \in P - \{p\}} \|p - p'\| \\ &\leq c \cdot \min_{p' \in P - \{p\}} \|p - p'\|. \end{aligned}$$

That shows $q \in N_c(p)$. Analogously it follows that $p \in N_c(q)$. □

Now we will show that the adaptive neighborhood graph cannot connect sample points from different connected components of M.

Lemma 3. *Let P be an ε-sample of M. The adaptive neighborhood graph $G_c(P)$ contains no edge that connects different connected components of M provided $\varepsilon < 1$ and $c < (1-\varepsilon)/2\varepsilon$.*

Proof. Let $p, q \in P$ be two sample points that are sampled from two different connected components of M. That is, the line segment pq connects two disjoint connected components of M. Assume $p \in M_i \in \mathcal{M}$ and consider the following continuous function on the pq:

$$g\colon\ pq \to \mathbb{R},\ x \mapsto \operatorname{dist}(x, M_i) - \operatorname{dist}(x, M \backslash M_i).$$

By construction we have $g(p) < 0$ and $g(q) > 0$. The continuity of g implies that there exists at least one point $x' \in pq$ with $g(x') = 0$. The point x' has to be a point of the medial axis of M. Hence we have

$$f(p),\ f(q) \leq \|p - q\|.$$

Let p' be the nearest neighbor of p in $P \backslash \{p\}$. From our assumption on c and Lemma 2(i) we get

$$c\|p - p'\| \leq \frac{2c\varepsilon}{1-\varepsilon} f(p) < f(p) \leq \|p - q\|.$$

By an analogous reasoning we also have, for the nearest neighbor q' of q in $P \backslash \{q\}$,

$$c\|q - q'\| \leq \frac{2c\varepsilon}{1-\varepsilon} f(q) < f(q) \leq \|p - q\|.$$

That is, the line segment pq cannot be an edge of any adaptive neighborhood graph $G_c(P)$ for $\varepsilon < 1$ and $c < (1-\varepsilon)/2\varepsilon$. □

Now we are ready to prove that the adaptive neighborhood graph has to be connected for each subset of sample points that belong to one connected component of M.

Lemma 4. *Let P be an (ε, δ)-sample of M and let $G_c[P \cap M_i]$ be the subgraph of the adaptive neighborhood graph $G_c(P)$ induced by $P \cap M_i$. The graph $G_c[P \cap M_i]$ is connected for every connected component $M_i \in \mathcal{M}$ of M provided $\varepsilon < 1$ and*

$$\frac{2\varepsilon}{(1-\varepsilon)\delta} \leq c \leq \frac{1-\varepsilon}{2\varepsilon}.$$

Proof. The restricted Delaunay graph on P is the intersection graph of the restricted Voronoi cells of the points in P. Let G be the induced subgraph on $P \cap M_i$ of the restricted Delaunay graph. It suffices to show that G is connected. The proof then follows from Corollary 1.

Let G' be a connected component of G and assume that $G' \neq G$. Let $P' \subset P$ be the vertex set of G' and let $\mathcal{F}$ be the set of all Voronoi facets shared by Voronoi cells of points from P' and $P \backslash P'$. By construction the set $\mathcal{F}$ is not empty and the intersection

$$\bigcup_{f \in \mathcal{F}} f \cap M_i$$

cannot be empty either since the collection of all Voronoi cells of points in P' do not cover the whole of M_i. That is, there exists a point $x \in M_i$ that is contained in the intersection $V_p \cap V_q$ where V_p is the Voronoi cell of a point $p \in P'$ and V_q is the Voronoi cell of a point $q \in P \backslash P'$. We know from Lemma 3 that q has to be a sample point of M_i, i.e., a vertex in the complement of G' in G. Hence the straight edge pq has to be a restricted Delaunay edge connecting a vertex p in G' with a vertex q in the complement of G' in G. That is a contradiction. □

We summarize our findings in the following theorem.

Theorem 1. *Let P be an (ε, δ)-sample of M. The adaptive neighborhood graph $G_c(P)$ has the same connectivity as M, provided $\varepsilon < 1$ and*

$$\frac{2\varepsilon}{(1-\varepsilon)\delta} \leq c \leq \frac{1-\varepsilon}{2\varepsilon}.$$

That is, two points $p, q \in P$ are connected by a path in M if and only if they are connected by a path in $G_c(P)$.

Note that the assumptions stated in the theorem are fulfilled if δ/ε is bounded from below by a suitable constant $\rho_0 > 0$ and ε is sufficiently small. For example $\varepsilon \leq \frac{1}{10}$, $\delta/\varepsilon \geq \rho_0 = \frac{1}{2}$, and $c = \frac{9}{2}$ will do.

4. Dimension Detection

In this section we present an algorithm which detects the local dimension of M at a sample point p, i.e., the dimension of the component of M that contains p. In contrast to reconstruction of the connectivity, dimension detection is a local task. We exploit this fact by considering only the set $N_c(p)$ instead of all neighbors of p in the adaptive neighborhood graph.

Here is a rough outline of our strategy: First we show that for a suitable choice of c and all sufficiently small $\varepsilon, \delta > 0$ such that δ/ε is bounded from below by a suitable constant, the following holds:

If P is an (ε, δ)-sample from M and if $p \in P$, then, on the one hand, all points $q \in N_c(P)$ lie very close to the tangent space T_pM of M at p, while, on the other hand, for any affine subspace L through p with $\dim L < \dim_p M$, there is some $q \in N_c(P)$ that is quite far away from L.

These findings give rise to an algorithm to determine the dimension of the manifold $M_i \in \mathcal{M}$ that contains p: Starting with $l = 1$, we apply an algorithm by Har-Peled and Varadarajan [15] to compute a 1.99-approximation L of the best-fit l-dimensional affine subspace through p for the set $N_c(P)$. We will see that for all suitable samples, the distance

$$\max_{q \in N_c(P)} \inf_{x \in L} \|x - q\|$$

is larger than some threshold if $l < k$ and it is at most $1.99/2$ times this threshold if $l \geq k$. Hence, the smallest l for which the maximum of the distances from L is smaller than the threshold gives us the dimension of the manifold $M_i \in \mathcal{M}$ that contains p.

We now make our strategy precise through a series of lemmas.

Lemma 5. *Let $p \in M$ and $u \in N_pM$, $\|u\| = 1$. Then the ball B of radius $f(p)$ centered at $p + f(p)u$ does not contain any points from M in its interior.*

Proof. For $\rho > 0$, consider the ball B_ρ of radius ρ centered at $p+\rho u$. We have $B_\rho \subseteq B_{\rho'}$ whenever $\rho \leq \rho'$, and since the balls B_ρ are tangent to M at p they do not contain points from M in their interior if ρ is sufficiently small. Let $r := \sup\{\rho > 0 : \mathrm{int} B_\rho \cap M = \emptyset\}$. Then $\mathrm{int} B_r \cap M = \emptyset$, and there is a point $q \in M$ such that $q \in \mathrm{bd} B_r$ and $q \in \mathrm{int} B_\rho$ for every $\rho > r$. In particular, $q \neq p$, so the center m of B_r belongs to the medial axis, and hence $r = \|m - p\| \geq f(p)$. □

Lemma 6 (Small Angle Lemma). *Let $p \in M$ and consider $q \in M$ with $\|p - q\| = tf(p)$, $0 < t < 1$. Then for every non-zero normal vector $u \in N_pM$, the distance from q to the hyperplane $p + u^\perp$ satisfies*

$$\mathrm{dist}(q, p + u^\perp) \leq \frac{t^2}{2} f(p). \tag{1}$$

In other words, the angle α between the segment pq and the tangent space T_pM satisfies

$$\sin \alpha \leq \frac{t}{2}.$$

Proof. If $q \in p + u^\perp$, we are done. Otherwise, there is a unique d-dimensional ball B which is tangent to $p + u^\perp$ at p and contains q on its boundary. Let m and r be the center

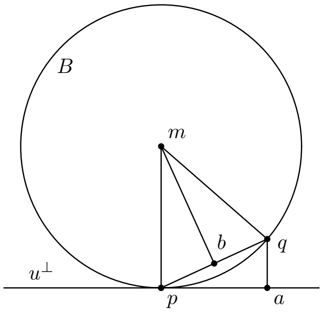

Fig. 1. The component of $q - p$ in any normal direction is small.

and the radius of B, respectively (if $\|u\| = 1$ and $\langle q - p, u\rangle > 0$, then $m = p + ru$). We have $r \geq f(p)$ (else q would be contained in the interior of the ball B' of radius $f(p)$ centered at $p + f(p)(u/\|u\|)$, contradicting Lemma 5). Let a be the orthogonal projection of q onto $p + u^{\perp}$, and let $b := \frac{1}{2}(p + q)$. Observe that the triangles pqa and mpb are similar (see Fig. 1), whence

$$\frac{\|q - a\|}{\|p - q\|} = \frac{\|p - q\|/2}{r}.$$

Therefore, $\|q - a\| = t^2 f(p)^2/2r \leq (t^2/2) f(p)$, as desired. Finally, for the last part of the assertion, either $q \in T_pM$, in which case $\alpha = 0$, or we can let a be the orthogonal projection of q onto T_pM and apply (1) with $u = q - a$. □

The previous lemma states that if we start from $p \in M$ and move by at most $tf(p)$ on the manifold, then there is a point on the tangent space that is very close, namely, at distance at most $t^2 f(p)$. We now consider the reverse of this statement: If we start from a point $p \in M$ and move by a sufficiently small amount t in direction of a unit tangent vector $v \in T_pM$, then there exists a point $q \in M$ very close to $p + tv$, namely, $\|p + tv - q\| = O(t^2)$. The following lemma specifies "sufficiently close" and the implicit constant in terms of the local feature size.

Lemma 7 (Close Point Lemma). *Let $p \in M$ and $v \in T_pM$, $\|v\| = 1$. For $0 < t < 1$, let $a(t) := p + tf(p) \cdot v$, and let $q(t)$ be the point on M closest to $a(t)$ (which is unique since $\|p - a(t)\| < f(p)$). If $t \leq t_0$ for some absolute constant t_0 ($t_0 = \frac{1}{4}$ works, for instance), then*

$$\|a(t) - q(t)\| < 2t^2 f(p). \tag{2}$$

We note that the constant factor of 2 for t^2 in (2) is somewhat arbitrary, any other constant strictly greater than $\frac{1}{2}$ will do if we adjust t_0 accordingly.

Proof. We proceed in two steps. We first prove that by general principles, (2) holds if t is sufficiently small, without being able to specify what exactly "sufficiently small" means. In the second step we show that if $t \leq t_0$ violates (2), then so does $t' := t/(1+t)$.

Therefore, if there existed some violator $t \leq t_0$, there would in fact be a whole sequence $t, t', t'', t''', \ldots$ of violators. Moreover, this sequence would converge to 0 (if $t \leq 1/n$, then $t' \leq 1/(n+1)$), eventually leading to a contradiction to what we proved in the first step.

1. Consider the geodesic γ through p on M in direction v. We assume that γ is parametrized by arc length and such that $\gamma(0) = p$. It follows that $\dot{\gamma}(0) = v$ and that $\ddot{\gamma}(0) \perp v$. Moreover, since γ is a geodesic, $\ddot{\gamma}(0) \perp w$ for all tangent vectors $w \in T_pM$ which are themselves orthogonal to v, see [16]. Therefore, $\ddot{\gamma}(0) \in N_pM$.

   By Taylor's formula, we have
$$\gamma(s) = \underbrace{\gamma(0)}_{p} + s\,\underbrace{\dot{\gamma}(0)}_{v} + \frac{s^2}{2}\ddot{\gamma}(0) + r(s), \tag{3}$$
   where $\|r(s)\|/s^2 \to 0$ (but we do not know how fast) as $s \to 0$. It follows that $\|\ddot{\gamma}(0)\| \leq 1/f(p)$, otherwise for sufficiently small s, $\gamma(s) \in M$ would contradict Lemma 6 (applied with $u = \ddot{\gamma}(0)$ and $t = s/f(p)$). Hence, if we choose t so small that $\|r(tf(p))\| < \frac{3}{2}t^2 f(p)$, say, then
$$\|a(t) - q(t)\| \leq \|p + tf(p) \cdot v - \gamma(tf(p))\| < 2t^2 f(p),$$
   which completes the first step.
2. Assume then that $t \leq t_0$ violates (2). We want to show that $t' := t/(1+t)$ is a violator as well. Without loss of generality, we may assume that $f(p) = 1$. For simplicity, we write $a := a(t)$, $q := q(t)$, and $a' := a(t')$. Observe that since q is the point on M closest to a, the (non-zero) vector $a - q$ is orthogonal to T_qM, and so is the unit vector $u := (a - q)/\|a - q\|$. Therefore, for any $r \leq f(q)$, the ball B of radius r centered at $m := q + ru$ does not contain any point from M in its interior. Thus, in order to prove that t' is again a violator, it is enough to show that for a suitable r to be specified below, the distance from a' to the boundary of B is at least $2(t')^2$, i.e.,
$$\|a' - m\|^2 \leq (r - 2(t')^2)^2. \tag{4}$$
   Let $x := \langle p - q, u\rangle$ and $y := \langle a - q, u\rangle$, see Fig. 2. We have $y = \|a - q\| \leq \|p - a\| = t$ (and also $y \geq 2t^2$, since t is a violator). Thus, as a first estimate, we get $\|p - q\| \leq \|p - a\| + \|a - q\| \leq 2t$, and so $f(q) \geq 1 - 2t$.

   Since u is orthogonal to T_qM and $\|p - q\| \leq 2t \leq 1 - 2t \leq f(q)$, we can apply Lemma 6 and conclude
$$|x| \leq \frac{\|p - q\|^2}{2f(q)}. \tag{5}$$
   Observe that the projection of $q - p$ onto $u^\perp$, which is the same as the projection of $a - p$, has length $\sqrt{t^2 - (y - x)^2}$. Therefore,
$$\|p - q\|^2 = x^2 + t^2 - (y - x)^2, \tag{6}$$

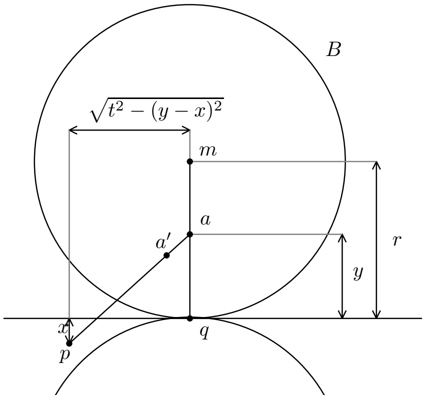

Fig. 2. t' is again a violator.

hence $x \le (t^2 - y^2 + 2xy)/2f(q)$, i.e., $x(1-y/f(q)) \le (t^2 - y^2)/2f(q)$. Since $y \le t < f(q)$, it follows that

$$x \le \frac{t^2 - y^2}{2f(q) - 2y}.$$

This is at most $y/2$ because $yf(q) \ge 2t^2(1-2t) \ge t^2$ for $t \le \frac{1}{4}$. Together with (6), $x \le y/2$ implies $\|p - q\| \le t$, so in fact $f(q) \ge 1 - t$. Therefore, we can set

$$r := 1 - t.$$

It remains to verify (4). By definition of t', we have $t - t' = tt'$, so we can write $a' - m = a - m - tt'v = a - m - t'(a - p)$. Therefore, by decomposing $a' - m$ into its components in direction u and in $u^\perp$, respectively, we obtain

$$\begin{aligned}\|a' - m\|^2 &= (r - y + t'(y - x))^2 + (t')^2(t^2 - (y - x)^2)\\ &= r^2 - 2ry + y^2 + 2t'(r - y)(y - x) + t^2(t')^2.\end{aligned}$$

Thus, in order to show (4), we have to prove

$$\begin{aligned}F(x, y, t, r) &:= 2ry - 4r(t')^2 - y^2 - 2t'(r - y)(y - x)\\ &\ge\ t^2(t')^2 - 4(t')^4.\end{aligned}$$

Observe that $(d/dx)F(x, y, t, r) = 2t'(r - y) \ge 0$, so we can substitute any lower bound for x without increasing F. By (5), we have

$$x \ge -\frac{t^2 - y^2}{2f(q) + 2y} \ge -\frac{t^2 - y^2}{2(1 - t) + 2y} \ge -\frac{y}{3},$$

since $t \le \frac{1}{4}$ and $y \ge 2t^2$, by assumption. Therefore, $F(x, y, t, r) \ge F(-y/3, y, t, r) = 2ry - 4r(t')^2 - y^2 - 2t'(r - y)(\frac{4}{3}y)$. Moreover, $(d/dy)F(-y/3, y, t, r) =$

$2r - 2y - \frac{8}{3}t'(r - 2y) > 2(r - y)(1 - \frac{8}{3}t') \geq 0$ since $y \leq r$ and $t \leq \frac{3}{5}$. Therefore, we can also replace y by the lower bound $2t^2$ and obtain

$$\begin{aligned} F(x, y, t, r) &\geq F(-\tfrac{2}{3}t^2, 2t^2, t, r) \\ &= 4rt^2 - 4r(t')^2 - 4t^4 - \tfrac{16}{3}t'(r - 2t^2)t^2 \\ &= \frac{4}{3}\frac{t^3(2 - 6t - t^2 + t^3)}{(1+t)^2} \\ &\geq \frac{t^4}{(1+t)^2} - 4\frac{t^4}{(1+t)^4}, \end{aligned}$$

as desired, provided that

$$\tfrac{1}{3}(8 + t - 50t^2 - 31t^3 + 4t^4 + 4t^5) \geq 0,$$

which is the case for $0 \leq t \leq 0.3712$. □

Note that we can apply the Small Angle Lemma to all $q \in N_c(p)$ if we choose $t = 2c \cdot \varepsilon/(1 - \varepsilon)$. That is, the lemma essentially states that the tangent space T_pM is a $k = \dim(T_pM)$-flat that approximates $N_c(p)$ very well. We will now show there are no lower-dimensional flats that well approximate $N_c(p)$.

Lemma 8 (Large Angle Lemma). *Let p be a sample point in $M_i \in \mathcal{M}$ and assume $\dim(M_i) = k$. There are absolute constants $\varepsilon_0 > 0$, $\rho_0 < 1$, and $c > 1$ satisfying*

$$\arcsin\left(\left(\frac{c\rho_0(\rho_0 - 2c\varepsilon_0)}{(\rho_0^2 + 2c^2\varepsilon_0)(\rho_0 - c\varepsilon_0)} - 1\right)^{-1}\right) < \frac{\pi}{4}$$

such that the following holds for every l-dimensional flat L through p with $l < k$:

Let P be an (ε, δ)-sample of M with $\varepsilon_0 \geq \varepsilon > \delta \geq \rho_0\varepsilon > 0$. Then the largest angle between L and any edge pq, $q \in N_c(p)$, is bounded from below by

$$\beta_0 := \frac{\pi}{4} - \arcsin\left(\left(\frac{c\rho_0(\rho_0 - 2c\varepsilon_0)}{(\rho_0^2 + 2c^2\varepsilon_0)(\rho_0 - c\varepsilon_0)} - 1\right)^{-1}\right).$$

Proof. We start the proof with a general construction. Let v be a unit vector in T_pM and let x be the following point in T_pM:

$$x := p + \left(\frac{c\varepsilon}{\rho_0} - \frac{\rho_0 - c\varepsilon}{\rho_0 - 2c\varepsilon}\varepsilon - \frac{2c^2\varepsilon^2}{\rho_0^2}\right) f(p) \cdot v.$$

From Lemma 7 we know that there exists a point $q \in M$ such that

$$\|x - q\| \leq 2\left(\frac{c\varepsilon}{\rho_0} - \frac{\rho_0 - c\varepsilon}{\rho_0 - 2c\varepsilon}\varepsilon - \frac{2c^2\varepsilon^2}{\rho_0^2}\right)^2 f(p) \leq \frac{2c^2\varepsilon^2}{\rho_0^2} f(p).$$

This implies

$$\|p - q\| \leq \|p - x\| + \|x - q\| \leq \frac{c\varepsilon}{\rho_0} f(p).$$

From the sampling condition we get that there exists a point $p' \in P$ such that

$$\|q - p'\| \le \varepsilon f(q) \le \frac{\varepsilon}{1 - (c\varepsilon/\rho_0)/(1 - c\varepsilon/\rho_0)} f(p) = \frac{\rho_0 - c\varepsilon}{\rho_0 - 2c\varepsilon} \varepsilon f(p).$$

Therefore,

$$\|x - p'\| \le \left(\frac{\rho_0 - c\varepsilon}{\rho_0 - 2c\varepsilon} \varepsilon + \frac{2c^2\varepsilon^2}{\rho_0^2} \right) f(p),$$

which in turn provides us with

$$\|p - p'\| \le \|p - x\| + \|x - p'\| \le \frac{c\varepsilon}{\rho_0} f(p) \le c\delta f(p).$$

That is, $p' \in N_c(p)$. We get for the inner angle α of the triangle xpp' at p that

$$\begin{aligned}
\sin\alpha \ &\le\ \frac{\|x - p'\|}{\|x - p\|} \le \frac{((\rho_0 - c\varepsilon)/(\rho_0 - 2c\varepsilon))\varepsilon + 2c^2\varepsilon^2/\rho_0^2}{c\varepsilon/\rho_0 - ((\rho_0 - c\varepsilon)/(\rho_0 - 2c\varepsilon))\varepsilon - 2c^2\varepsilon^2/\rho_0^2} \\
&= \left(\frac{c\rho_0(\rho_0 - 2c\varepsilon)}{\rho_0^2(\rho_0 - c\varepsilon) + 2c^2\varepsilon(\rho_0 - 2c\varepsilon)} - 1 \right)^{-1} \\
&\le \left(\frac{c\rho_0(\rho_0 - 2c\varepsilon)}{(\rho_0^2 + 2c^2\varepsilon)(\rho_0 - c\varepsilon)} - 1 \right)^{-1} \\
&\le \left(\frac{c\rho_0(\rho_0 - 2c\varepsilon_0)}{(\rho_0^2 + 2c^2\varepsilon_0)(\rho_0 - c\varepsilon_0)} - 1 \right)^{-1}.
\end{aligned}$$

Note that this bound on $\sin\alpha$ goes to $\rho_0/(c - \rho_0)$ as ε_0 goes to 0.

Now we are prepared to prove the lemma. We distinguish two cases. Either there exists a point $y \in L$ such that the vector $w := y - p$ makes an angle β larger than $\pi/4$ with its projection on T_pM or all such vectors make an angle less than or equal to $\pi/4$ with their projection on T_pM.

In the first case let $\pi(w)$ be the unit vector in the direction of the projection of w on T_pM. If we set v in the above calculations to be $\pi(w)$, then the inner angle φ of the triangle ypp' at p can be bounded from below by using the triangle inequality for angles as follows:

$$\varphi \ge \beta - \alpha > \frac{\pi}{4} - \alpha.$$

In the second case let L' be the projection of L onto T_pM. Since $l < k$ we find a unit vector v in T_pM that is orthogonal to L'. With this vector v we find as above $p' \in P$. Let $y \ne p$ be some point in L and let y' be its projection on T_pM. Let β be the smaller of the two angles made by $y - p$ and $y' - p$. From our assumption we have $\beta \le \pi/4$. We can bound the inner angle φ of the triangle ypp' at p from below by using the triangle inequality for angles twice:

$$\varphi \ge \frac{\pi}{2} - \beta - \alpha \ge \frac{\pi}{4} - \alpha.$$

This proves our claim. □

We summarize:

Theorem 2. *There are absolute constants* $\varepsilon_0 > 0$, $\rho_0 > 1$, *and* $c \geq 1$ *such that for all* $\varepsilon_0 > \varepsilon > \delta \geq \rho_0 \varepsilon > 0$, *the following holds*:

Suppose P *is an* (ε, δ)*-sample from* M *and* $p \in P$ *with* $\dim_p M = k$ (*i.e.*, $p \in M_i$ *with* $\dim M_i = k$). *Then*

1. *the maximal angle* α *between the tangent space* $T_p M$ *and any edge* pq, $q \in N_c(p)$, *satisfies* $\sin \alpha \leq c\varepsilon/(1-\varepsilon)$, *but*
2. *for any affine subspace* L *through* p *of dimension* $\dim L < k$, *there is a point* $q \in N_c(p)$ *such that the angle* β *between* pq *and* L *is at least* β_0, *as defined in Lemma* 8.

That is, $\alpha << \beta$ *if* ε_0 *is sufficiently small.*

We use Theorem 2 to devise an algorithm for dimension detection. The running time of the algorithm will depend on the size of $N_c(p)$, so we make a brief detour to bound the latter.

Lemma 9. *Suppose that* ε, δ, *and* P *are as in Theorem* 2. *Assume that*

$$\frac{2c^2\varepsilon_0}{(1-\varepsilon_0)(1-(2c+1)\varepsilon_0)} < \rho_0.$$

Then, for $p \in P$,

$$|N_c(p)| = 2^{O(k)},$$

where $k = \dim_p M$, *with the constant of proportionality depending on* ε_0, ρ_0, *and* c.

Proof. From Lemma 2 we know that any point $q \in N_c(p)$ has distance at most $(2c\varepsilon/(1-\varepsilon)) f(p)$ from p. We set

$$\eta := \frac{2c\varepsilon}{1-\varepsilon} \quad \text{and} \quad \eta_0 := \frac{2c\varepsilon_0}{1-\varepsilon_0}.$$

Observe that $\eta \leq \eta_0$. Our assumption guarantees that we have that $\delta f(q) \geq \delta(1-\eta) f(p) > (\eta/2) f(p) \geq \operatorname{dist}(q, T_p M)$, so the d-dimensional ball of radius $\delta f(q)$ centered at $q \in N_c(p)$ intersects $T_p M$ in a k-dimensional ball centered at the projection of q onto $T_p M$. By Pythagoras' Theorem, the squared radius of that k-dimensional ball is

$$\begin{aligned} \delta^2 f(q)^2 - d(q, T_p M)^2 &\geq \delta^2 (1-\eta)^2 f(p)^2 - \eta^4 f(p)/4 \\ &= (\delta^2(1-\eta)^2 - \eta^4/4) f(p)^2 =: r^2. \end{aligned}$$

Therefore, by our sampling condition, the k-dimensional balls of radius r centered at the projections of the points from $N_c(p)$ onto $T_p M$ are disjoint, and they are all contained in the k-dimensional ball of radius $\eta f(p) + r$ centered at p. Since $r < \delta f(p) \leq$

$(2\varepsilon/(1-\varepsilon))f(p)$, we have $\eta f(p) + r \le (2(c+1)\varepsilon/(1-\varepsilon))f(p)$, and so $|N_c(p)|$ is at most

$$\begin{aligned}\frac{(\eta f(p)+r)^k}{r^k} &\le \left(\frac{16(c+1)^2\varepsilon^2/(1-\varepsilon)^2}{4\delta^2(1-\eta)^2-\eta^4}\right)^{k/2}\\ &\le \left(\frac{16(c+1)^2\varepsilon^2/(1-\varepsilon)^2}{4\varepsilon^2(1-\eta)^2\rho_0^2-\eta^4}\right)^{k/2}\\ &= \left(\frac{4(c+1)^2/(1-\varepsilon)^2}{(1-\eta)^2\rho_0^2-4c^4\varepsilon^2/(1-\varepsilon)^4}\right)^{k/2}\\ &\le \left(\frac{4(c+1)^2/(1-\varepsilon_0)^2}{(1-\eta_0)^2\rho_0^2-4c^4\varepsilon_0^2/(1-\varepsilon_0)^4}\right)^{k/2}\\ &= 2^{O(k)}. \qquad \square\end{aligned}$$

For dimension detection we use, with $\xi = 0.99$ say, the following result of Har-Peled and Varadarajan [15]: Let $U \subset \mathbb{R}^d$ be a finite set of points. Then, for $0 < \xi < 1$ and any integer $0 \le l \le d-1$, one can compute a $(1+\xi)$-approximation L for the best-fit l-dimensional linear subspace for U in time

$$d|U|^{O(l^6/\xi^5 \log(l/\xi))}. \tag{7}$$

That is, one can compute an l-dimensional linear subspace L such that

$$\max_{u\in U} \operatorname{dist}(u, L) \le (1+\xi) \min_{L'} \max_{u\in U} \operatorname{dist}(u, L'),$$

where dist denotes the orthogonal distance and the minimum is taken over all l-dimensional linear subspaces L' (i.e., l-flats which contain the origin).

Theorem 3. *There are constants ε_0, ρ_0, and c such that, for $\varepsilon_0 > \varepsilon > \delta \ge \rho_0\varepsilon > 0$, the following holds*:

Suppose we are given a finite set P of n points in $\mathbb{R}^d$ and we are guaranteed that it is an (ε, δ)-sample for M. Then, for each $p \in P$, if we have precomputed $N_c(p)$, the local dimension $k = \dim_p M$ at p can be computed in time

$$d2^{O(k^7 \log k)}.$$

The actual computation of the neighborhoods $N_c(p)$ can be trivially done in linear time for a single point, and hence in time $O(n^2)$ for the entire point set (for high codimensions, this will still be much faster than computing the Delaunay triangulation). It seems natural, at least if the ambient dimension d is considered fixed, to try to speed up this computation by using approximate proximity data structures as in, for instance, [5], but we do not pursue this issue further in this paper.

Proof. Choose ε_0, ρ_0, and c according to Theorem 2 such that

$$\frac{c\varepsilon_0}{1-\varepsilon_0} \le \frac{\sin\beta_0}{2}.$$

To compute the local dimension at $p \in P$, we proceed as follows:

1. Compute $N_c(p)$ and re-normalize its elements to obtain a set of unit vectors
$$U = U_c(p) = \left\{ \frac{q-p}{\|q-p\|} : q \in N_c(p) \right\} \subset \mathbb{R}^d.$$
2. Starting with $l = 1$, compute a 1.99-approximation L for the best-fit l-dimensional linear subspace for U. Stop and output $\dim_p M = l$ as soon as
$$\max_{u \in U} \operatorname{dist}(u, L) \leq \frac{1.99}{2} \sin(\beta_0).$$

The correctness of this algorithm is immediate from Theorem 2, and the running time follows from (7) and the fact that $|N_c(p)| \leq 2^{O(k)}$, as proved in Lemma 9. □

5. Approximation of Geodesic Distances

For $x, y \in M$, we denote by $\operatorname{dist}_M(x, y)$ the *geodesic distance* between x and y in M. If p and q lie in different connected components of M, then this distance is set to be ∞; otherwise it is defined as the infimum of the lengths $L(\gamma)$ over all rectifiable continuous curves $\gamma\colon [0, 1] \to M$ connecting p and q, i.e., $\gamma(0) = p$ and $\gamma(1) = q$. Recall that the length $L(\gamma)$ is defined as the supremum of $\sum_{i=1}^{N} \|\gamma(t_i) - \gamma(t_{i-1})\|$ over all finite subdivisions $0 = t_0 < t_1 < \cdots < t_N = 1$ of the parameter interval, and that γ is called rectifiable if $L(\gamma) < \infty$.

Since each connected component M_i of M is a smooth compact manifold, for any two points $p, q \in M_i$ there exists a shortest geodesic γ connecting p and q such that $L(\gamma) = \operatorname{dist}_M(p, q)$.

Further, if G is the adaptive neighborhood graph for P and $p, q \in P$, let $\operatorname{dist}_G(p, q)$ be shortest-path distance between p and q in the geometric graph G, i.e., the minimum of $\sum_{i=1}^{m} \|p_i - p_{i-1}\|$ over all paths $p = p_0, p_1, \ldots, p_m = q$ between p and q in G.

For suitable values of the constants c, ρ_0, and ε_0, the distances in the adaptive neighborhood graph are good approximations for the geodesic distances. This is made precise in Theorems 4 and 5 following. It follows that by applying Dijkstra's algorithm to the adaptive neighborhood graph $G_c(P)$, we can very efficiently approximate geodesic distances in M.

Since we know from Theorem 1 that G and M have the same connected components, it suffices to consider the case that M is connected, and we assume so throughout this section.

Theorem 4. *There are constants c, ρ, and ε_0 such that, for $\varepsilon_0 \geq \varepsilon > \delta \geq \rho_0\varepsilon > 0$, the following holds*:

If G is the c-adaptive neighborhood graph for an (ε, δ)-sample P from M, then for all $p, q \in P$,
$$\operatorname{dist}_M(p, q) \leq (1 + O(\varepsilon^2))\operatorname{dist}_G(p, q)$$
as $\varepsilon \to 0$.

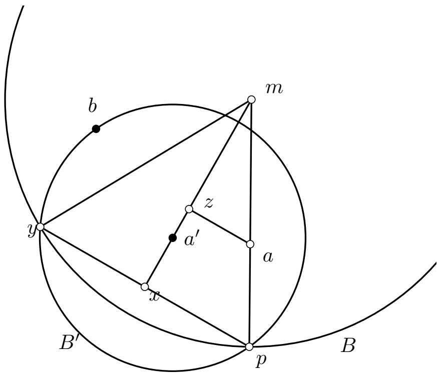

Fig. 3. Estimating the diameter of $B' \setminus \text{int}B$.

The proof of the theorem proceeds in a somewhat roundabout way. We first establish the following technical lemma:

Lemma 10. *Let $a, a' \in \mathbb{R}^d$, and let p and p' be the points of M closest to a and a', respectively. Assume that $\|a - a'\| \le \|p - a\| \le f(p)/2$. Then*

$$\|p - p'\| \le \frac{2f(p)}{f(p) - \|p - a\|}\|a - a'\| \le 4\|a - a'\|.$$

Proof. Since p is the point in M closest to a, we have $a - p \in N_pM$. We may assume that $p \neq a$, otherwise the assertion of the lemma is trivial.

Consider then the ball B of radius $f(p)$ centered at $m := p + f(p) \cdot (p - a)/\|p - a\|$. This ball is tangent to M at p, and M does not intersect the interior int$B$.

Moreover, let B' be the ball of radius $\|a' - p\|$ centered at a', see Fig. 3. Since p' is the point of M that is closest to a', we have $p' \in B' \setminus \text{int}B$.

Now we estimate the diameter of $B' \setminus \text{int}B$. Let $b := 2a' - p$ be the point antipodal to p in B'. We claim that $b \in B$.

Suppose we had already shown this. The boundaries of B and B' intersect in a $(d - 2)$-dimensional sphere S. Let x be the center of that sphere. Since $b \in B$, the point in $B' \setminus \text{int}B$ that is farthest from p is the point $y := 2x - p$ that is antipodal to p in S. Hence, $\|p - p'\| \le \|p - y\| = 2\|p - x\|$, and $\|p - x\|/\|p - m\| = \|z - a\|/\|a - m\|$, where z is the orthogonal projection of a onto the line through m and a'. This establishes the lemma because $\|p - m\| = f(p)$ and $\|z - a\| \le \|a - a'\|$, so it suffices to prove the claim.

In order to see that $b \in B$, first observe that our assumption $\|a - a'\| \le \|p - a\| \le f(p)/2$ implies that a' is contained in the ball B'' with diameter $\|m - p\|$ through m and p, see Fig. 4. Therefore, by Thales' theorem, the angle $\alpha = \angle pa'm$ is at least $\pi/2$. It follows that the angle $\beta = \angle ma'b = \pi - \alpha$ satisfies $\beta \le \alpha$.

Finally, by the Cosine Theorem, we have

$$\|b - m\|^2 = \|b - a'\|^2 + \|m - a'\|^2 - 2\|b - a'\| \cdot \|m - a'\| \cos\beta$$

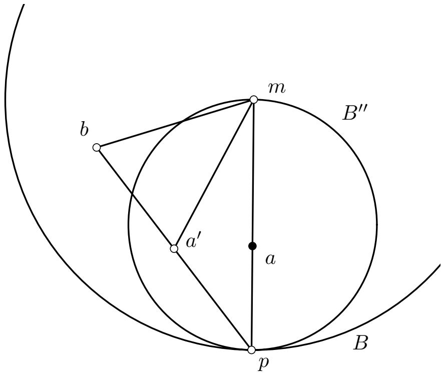

Fig. 4. Proving that $b \in B$.

and

$$\|p - m\|^2 = \|m - a'\|^2 + \|p - a'\|^2 - 2\|m - a'\| \cdot \|p - a'\| \cos \alpha.$$

Since $\|b - a'\| = \|p - a'\|$ and $\alpha \geq \beta$, we conclude $\|b - m\| \leq \|p - m\| = f(p)$, i.e., $b \in B$. □

We also need the following result due to Schmidt [18] (see also [1] for a more general version).

Theorem (Schmidt). Let $\gamma\colon [0, L] \to \mathbb{R}^d$ be a C^2 curve without self-intersections. We assume that γ is parametrized by arclength, i.e., that $\|\dot{\gamma}(t)\| = 1$ for all t. In particular, L is the length of γ, and $L = L(\gamma) = \int_0^L \|\dot{\gamma}(t)\|\, dt$. If the *total curvature* $C(\gamma) = \int_0^L \|\ddot{\gamma}(t)\|\, dt$ is less than π, then the distance $\|\gamma(0) - \gamma(L)\|$ between the two endpoints of γ is at least $L(\gamma) \cdot \cos(C(\gamma)/2)$.

Proof of Theorem 4. We assume that ε, δ, and c satisfy the assumptions of Theorem 1. If $p = p_0, p_1, \ldots, p_N = q$ form a shortest path in $G = G_c(P)$ connecting p and q, i.e., $\mathrm{dist}_G(p, q) = \sum_{i=1}^N \|p_i - p_{i-1}\|$, then $\mathrm{dist}_M(p, q) \leq \sum_{i=1}^N \mathrm{dist}_M(p_i, p_{i-1})$, so it suffices to prove the theorem for the case that p and q are adjacent in G.

By Lemma 2, we can write $\|p - q\| = \eta f(p)$, with $\eta \leq 2c\varepsilon/(1 - \varepsilon)$. We assume that ε and c are chosen so that $\eta < \frac{1}{4}$ and $4\eta/(1 - 4\eta) < \pi$.

As a first step, we show that there exists a curve β in M connecting p and q such that $L(\beta) \leq 4\|p - q\|$.

We define $\beta\colon [0, 1] \to M$ as follows: For $t \in [0, 1]$, let $a(t) := p + t(q - p)$, and let $\beta(t)$ be the point in M closest to $a(t)$ (which is unique since $\|a(t) - p\| < f(p)$). It is not hard to verify that this defines indeed a continuous curve (we will not need any stronger smoothness properties of β). Further, we have $\|\beta(t) - p\| \leq 2\eta f(p)$, hence $f(\beta(t)) \geq (1 - 2\eta) f(p)$ for all t. Now, consider a subdivision $0 = t_0 < t_1 < \cdots < t_N = 1$ of the interval $[0, 1]$. If the subdivision is sufficiently fine, then any two consecutive points

$a(t_i), a(t_{i-1})$ satisfy the assumptions of Lemma 10. Therefore,

$$\sum_{i=1}^{N} \|\beta(t_i) - \beta(t_{i-1})\| \leq 4 \sum_{i=1}^{N} \|a(t_i) - a(t_{i-1})\| = 4\|p - q\|,$$

and since this holds for all subdivisions, we conclude that $L(\beta) \leq 4\|p - q\|$.

Now, let γ be a shortest geodesic connecting p and q in M. We assume that γ is parametrized by arc length. We write $L(\gamma) = x\eta f(p)$, and our aim is to show that $x = 1 + O(\varepsilon^2)$. As a first estimate, we have $L(\gamma) \leq L(\beta) \leq 4\|p - q\|$, i.e., $x \leq 4$. In particular, $\|\gamma(t) - p\| \leq 4\|p - q\| = 4\eta f(p)$ for all $t \in [0, L(\gamma)]$.

It follows that $f(\gamma(t)) \geq (1 - 4\eta) f(p)$ for all t. Hence, $\|\ddot{\gamma}(t)\| \leq 1/(1 - 4\eta) f(p)$, as we saw in the proof of the Close Point Lemma (Lemma 7), and therefore

$$C(\gamma) = \int_0^{L(\gamma)} \|\ddot{\gamma}(t)\|\, dt \leq \frac{L(\gamma)}{(1 - 4\eta) f(p)} = \frac{x\eta}{1 - 4\eta} < \pi.$$

We can now apply Schmidt's theorem, which yields

$$\eta f(p) = \|p - q\| \geq L(\gamma) \cos\left(\frac{C(\gamma)}{2}\right) \geq x\eta f(p) \left(1 - O\left(\left(\frac{x\eta}{2(1 - 4\eta)}\right)^2\right)\right),$$

since $\cos(t) = 1 - O(t^2)$ for small t. We also know that $x \leq 4$, so it follows that $1 \geq x(1 - O(\eta^2))$, hence $x = 1 + O(\eta^2) = 1 + O(\varepsilon^2)$, as desired. □

Thus, we obtain better and better upper estimates for the geodesic distances as the sample becomes denser. On the other hand, for a lower bound, we can only guarantee a constant depending on c:

Theorem 5. *Fix ρ_0. For all $c \geq 1$, there exists ε_0 such that for all $\varepsilon_0 \geq \varepsilon > \delta \geq \rho_0\varepsilon > 0$, we have*

$$\mathrm{dist}_G(p, q) \leq \left(1 + O\left(\frac{1}{c}\right)\right) \mathrm{dist}_M(p, q)$$

for all $p, q \in P$.

Proof. Let γ be a geodesic of length $\mathrm{dist}_M(p, q)$ connecting p and q. For every point $x \in \gamma$ let $q(x)$ be the point in P closest to x. We set

$$\eta := \frac{1}{\rho_0(1 - \varepsilon)}.$$

We construct a finite sequence of points $p = x_1, x_2, \ldots$ on γ. We set $m_i := \max\{f(q(x_{i-1})), f(q(x_{x_i}))\}$ for $i > 1$. Let x_2 be the point on γ farthest from $x_1 = p = q(x_1)$ in the order along γ such that

$$\mathrm{dist}_M(x_1, x_2) = (c - \eta)\rho_0\varepsilon m_2.$$

If such a point x_2 does not exist, then

$$\begin{aligned}\|p-q\| &\leq (c-\eta)\rho_0\varepsilon\max\{f(p),\, f(q)\}\\ &\leq c\rho_0\varepsilon\max\{f(p),\, f(q)\}\\ &\leq c\delta\max\{f(p),\, f(q)\}.\end{aligned}$$

Thus pq is an edge of $G_c(P)$ and the claim follows immediately. Otherwise we have from the sampling condition

$$\|x_2-q(x_2)\| \leq \varepsilon f(x_2) \quad \text{thus} \quad \|x_2-q(x_2)\| \leq \frac{\varepsilon}{1-\varepsilon} f(q(x_2)).$$

We derive

$$\begin{aligned}\|x_1-q(x_2)\| &\leq \|x_1-x_2\| + \|x_2-q(x_2)\|\\ &\leq \mathrm{dist}_M(x_1,x_2) + \frac{\varepsilon}{1-\varepsilon} f(q(x_2))\\ &= (c-\eta)\rho_0\varepsilon m_2 + \frac{\varepsilon}{1-\varepsilon} f(q(x_2))\\ &\leq (c-\eta)\rho_0\varepsilon m_2 + \frac{\varepsilon}{1-\varepsilon} m_2\\ &\leq \left(c-\eta+\frac{1}{\rho_0(1-\varepsilon)}\right)\rho_0\varepsilon m_2\\ &= c\rho_0\varepsilon m_2 \leq c\delta m_2.\end{aligned}$$

Hence $x_1q(x_2)$ is an edge of $G_c(P)$. That is we can approximate $\mathrm{dist}_M(x_1,x_2)$ by $\|x_1-q(x_2)\|$ in $G_c(P)$. We get for the approximation quality

$$\begin{aligned}\frac{\|x_1-q(x_2)\|}{\mathrm{dist}_M(x_1,x_2)} &\leq \frac{c\rho_0\varepsilon m_2}{(c-\eta)\rho_0\varepsilon m_2} = \frac{c}{c-\eta}\\ &= 1+\frac{\eta}{c-\eta} = 1+O\left(\frac{1}{c}\right).\end{aligned}$$

We proceed on γ by choosing x_3 to be the point on γ farthest from x_2 such that

$$\mathrm{dist}_M(x_2,x_3) = (c-2\eta)\rho_0\varepsilon\max\{f(q(x_2)),\, f(q(x_3))\}.$$

If such a point does not exist, then we stop with x_2. Otherwise we get from a calculation similar to the one above that $q(x_2)q(x_3)$ is an edge in $G_c(P)$. That is, we can approximate $\mathrm{dist}_M(x_2,x_3)$ by $\|q(x_2)-q(x_3)\|$ in $G_c(P)$. We get for the approximation quality

$$\frac{\|q(x_2)-q(x_3)\|}{\mathrm{dist}_M(x_2,x_3)} \leq 1+\frac{2\eta}{c-2\eta} = 1+O\left(\frac{1}{c}\right).$$

We continue this construction with $x_4, x_5, \ldots$. Since γ has finite length this sequence has to be finite. Let x_n be the last point of the sequence. We have

$$\mathrm{dist}_M(x_n, q) < (c - 2\eta)\rho_0\varepsilon \max\{f(q(x_n)), f(q)\}.$$

By continuity we can find $c/2 < c' \leq c$ and a sequence $p = y_1, \ldots, y_n, y_{n+1} = q$ constructed the same way as the sequence $p = x_1, \ldots, x_n$ only replacing c by c' such that for all $i = 1, \ldots, n$ it holds that

$$\mathrm{dist}_M(y_i, y_{i+1}) < (c' - 2\eta)\rho_0\varepsilon \max\{f(q(y_i)), f(q(y_{i+1}))\}.$$

The length of the path $q(y_1)q(y_2)\cdots q_(y_{n+1})$ in $G_c(P)$ approximates the length of γ up to a factor of $1 + O(1/c') = 1 + O(1/c)$. □

Here is an example that shows that the dependence on c of the previous estimate is unavoidable, i.e., that the lower estimate of the geodesic distance in terms of the graph distance need not become better as the density of the sample increases.

For the sake of concreteness, let $c = 4$ and $\rho_0 = \frac{1}{2}$. Let M consist of two large parallel squares at distance 2 in 3-space; to compactify the example, we join the boundaries of the squares by four half-cylinders of radius 1 (for the sides) and four quarter-spheres for radius 1 (for the corners). Thus, $f(x) = 1$ for all $x \in M$.

For $\varepsilon > 0$, consider a regular hexagonal grid of edge length $\delta = \varepsilon/2$ embedded on the flat portion of M (one grid on each of the squares, but we only work with one copy), and let the sample P consist of the subset of the vertices of the grid as shown in Fig. 5, extended to a uniform sample on the non-flat parts of M in your favorite fashion.

For the points p and q indicated, we have $\mathrm{dist}_M(p, q) = 6\delta$ and $\mathrm{dist}_G(p, q) = 2\sqrt{13}\delta$ for all $\delta = \varepsilon/2 > 0$.

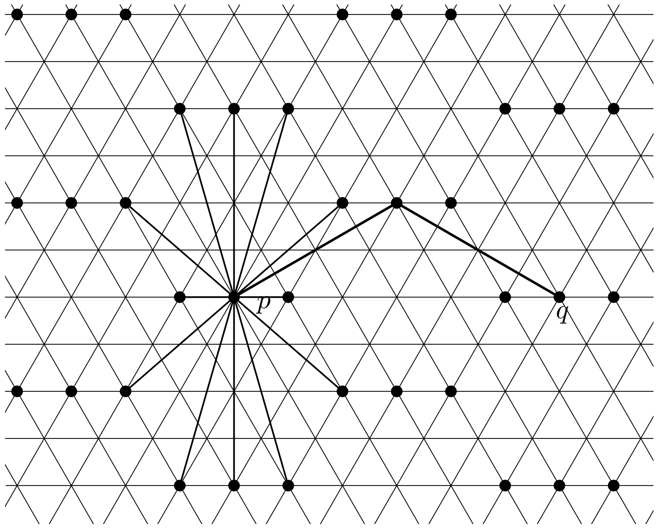

Fig. 5. A hexagonal grid with the neighborhood of p and the shortest path from p to q.

6. Conclusion

We have shown that the adaptive neighborhood graph can replace the Delaunay triangulation for some tasks in sample-based modeling. That is important since the Delaunay triangulation is prohibitive to compute in high dimensions. With the adaptive neighborhood graph it becomes feasible to provably correctly solve problems as inferring the dimension and connectivity of a manifold from a sample even if the ambient dimension is very high.

The (ε, δ)-sampling condition is quite strict and it is reasonable to assume that it is hardly ever met in practice. Nevertheless, a combination of the adaptive neighborhood graph with the k nearest neighbor approach, i.e., building the neighborhood graph on k nearest neighbors instead of just the nearest neighbor, should remove on practical data sets, e.g., locally uniform random samples, the disadvantages of both approaches when applied alone (too strict sampling condition for the neighborhood graph and non-adaptivity to the dimension of the k nearest neighbors).

We are quite confident that the adaptive neighborhood graph will be useful even for the more general problem of manifold reconstruction.

References

1. A.D. Aleksandrov and Y.G. Reshetnyak, *General Theory of Irregular Curves*. Kluwer Academic, Dordrecht (1989).
2. N. Amenta and M. Bern. Surface reconstruction by Voronoi filtering. *Discrete Comput. Geom.*, **22** (1999), 481–504.
3. N. Amenta, S. Choi, T.K. Dey, and N. Leekha. A simple algorithm for homeomorphic surface reconstruction. In *Proc.* 16*th. ACM Sympos. Comput. Geom.*, pp. 213–222 (2000).
4. N. Amenta, S. Choi, and R.K. Kolluri. The power crust, unions of balls, and the medial axis transform. *Comp. Geom. Theory Appl.*, **19**(2–3) (2001), 127–153.
5. S. Arya, T. Malamatos, and D.M. Mount. Space-efficient approximate Voronoi diagrams. In *Proc.* 34*th ACM Sympos. Theory Comput.*, pp. 721–730 (2002).
6. D. Attali, J.-D. Boissonnat, and A. Lieutier. Complexity of the Delaunay triangulation of points on surfaces: the smooth case. In *Proc.* 19*th ACM Sympos. Comput. Geom.*, pp. 201–210 (2003).
7. M. Bern et al. Emerging Challenges in Computational Topology. NSF report (1999).
8. M. Bernstein, V. de Silva, J.C. Langford and J. Tenenbaum. Graph approximations to geodesics on embedded manifolds. Manuscript (2000).
9. C. Bregler and S.M. Omohundro. Nonlinear manifold learning for visual speech recognition. In *Proc. Int. Conf. Comput. Vision*, pp. 494–499 (1995).
10. C. Bregler and S.M. Omohundro. Nonlinear image interpolation using manifold learning. In G. Tesauro, O.S. Touretzky, and T.K. Leen (eds.), *Advances in Neural Information Processing Systems*, vol. 7. MIT Press, Cambridge, MA, pp. 973–980 (1995).
11. T.K. Dey and J. Giesen. Detecting undersampling in surface reconstruction. In *Proc.* 17*th. ACM Sympos. Comput. Geom.*, pp. 257–263 (2001).
12. T.K. Dey, J. Giesen, S. Goswami and W. Zhao. Shape dimension and approximation from samples. In *Proc.* 13*th ACM–SIAM Sympos. Discrete Algorithms*, pp. 772–780 (2002).
13. J. Erickson. Dense point sets have sparse Delaunay triangulations. *Discrete Comput. Geom.*, to appear.
14. S. Funke and E. Ramos. Smooth-surface reconstruction in near-linear time. In *Proc.* 13*th ACM–SIAM Sympos. Discrete Algorithms*, pp. 781–790 (2002).
15. S. Har-Peled and K. Varadarajan. Projective clustering in high dimensions using core-sets. In *Proc.* 18*th Sympos. Comp. Geom.*, pp. 312–318 (2002).

16. J. Jost. *Differentialgeometrie und Minimalflächen*. Springer-Verlag, Berlin (1991).
17. T. Martinetz and K. Schulten. Topology representing networks. *Neural Networks*, **7** (1994), 507–522.
18. E. Schmidt. Über das Extremum der Bogenlänge einer Raumkurve bei vorgeschriebenen Einschränkungen ihrer Krümmung. *Sitzungsber. Preuss. Akad. Wiss. Berlin Phys. Math. Kl.*, **25** (1925), 485–490.
19. J. Tenenbaum, V. de Silva and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, **290** (2000), 2319–2322.

Received May 15, 2003, *and in revised form January* 15, 2004. *Online publication June* 1, 2004.