

# Dimension Detection via Slivers\*

Siu-Wing Cheng<sup>†</sup>

Man-Kwun Chiu<sup>†</sup>

## Abstract

We present a novel approach to estimate the dimension  $m$  of an unknown manifold  $M \subset \mathbb{R}^d$  with positive reach from a set of point samples  $P \subset M$ . It works by analyzing the shape of simplices formed by point samples. Suppose that  $P$  is drawn from  $M$  according to a Poisson process with an unknown parameter  $\lambda$ . Let  $k$  be some fixed positive integer. When  $\lambda$  is large enough, we prove that the dimension can be correctly output in  $O(kd|P|^{1+1/k})$  time with probability greater than  $1 - 2^{-k}$ . We experimented with a practical variant and showed that its performance is competitive with several previous methods.

## 1 Introduction.

A lot of data have very high extrinsic dimension. For example, a collection of  $64 \times 64$  black and white images can be viewed as a point set  $P$  in 4096-dimensional space. When the images are related, it is often postulated that  $P$  lives on a manifold  $M$  of much lower dimension. The dimension detection problem is to compute the dimension of  $M$  given a set  $P$  of point samples drawn from  $M$ . Knowing the manifold dimension helps reconstructing the manifold [2, 4] and embedding the manifold in the parameter space [1, 9, 19, 20].

In machine learning, there are near-neighbor methods and global methods for estimating the manifold dimension. A review of concepts and some earlier methods can be found in [7]. We mention the more recent works. A typical global method is to perform manifold embedding [1, 9, 19, 20] for different target dimensions and decide according to some criteria. This involves repeatedly solving an eigenvalue problem for a  $|P|$  by  $|P|$  matrix. The GMST method [6] is an exception and it tracks the lengths of a sequence of minimal spanning trees. Near-neighbor methods analyze some counts of point samples in a neighborhood [10, 13, 14, 15, 16, 18]. It is a research issue to determine the neighborhood size.

In computational geometry, Dey et al. [8] gave a provably correct algorithm by analyzing the shape of

cells in the Voronoi diagram of  $P$ . The running time is  $O(|P|^{\lceil d/2 \rceil})$ , where  $d$  is the ambient space dimension. Experimental results were reported for the case of  $d = 3$  only. Giesen and Wagner [11] found a neighborhood in  $O(d|P|)$  time and then determine the dimension  $m$  of  $M$  by fitting in  $O(d2^{O(m^7 \log m)})$  time the best affine subspace to the neighborhood. Implementation is not reported in [11]. Cheng et al. [5] proved that the dimension can be estimated by applying PCA to a neighborhood in  $O(d2^{O(m)})$  time after computing the neighborhood in  $O(d|P|)$  time. Extremely high sampling density is needed in the experiments in [5].

We present a novel approach to estimate the manifold dimension by analyzing the shape of simplices formed by point samples in a neighborhood. Our approach is based on detecting *slivers* [3, 4] which are simplices with negligible volume. Let  $k$  be some fixed positive integer. When  $P$  is drawn from  $M$  according to a Poisson process with an unknown parameter  $\lambda$  such that  $\lambda = \Omega(2^{\Theta(m^6)} + 2^{\Theta(km^2)})$ , we prove that the dimension  $m$  can be correctly output in  $O(kd|P|^{1+1/k})$  time with probability  $1 - 2^{-k}$ . Notice that  $|P| = \Theta(\lambda)$  with high probability due to the Poisson process. We experimented with a practical variant of our algorithm and demonstrated that its performance is competitive with several previous methods. Also, high sampling density is not needed; for example, 500 points suffice for  $S^5$ .

In comparison, the running time in [8] is exponential in  $d$ . Our running time is better than the running time of  $O(d2^{O(m^7 \log m)})$  in [11] as long as  $|P| = o(2^{O(m^7 \log m)})$ . The local PCA in [5] requires a lot more point samples than our program in the experiments. For example, 150K points are needed for  $S^5$  in [5] but 500 points already suffice for our program.

## 2 Notation.

For any point or vector  $x$  in  $\mathbb{R}^d$ , we use  $\|x\|$  to denote the Euclidean norm of  $x$ . Given two points  $x$  and  $y$ ,  $\|x - y\|$  is the distance between them. For any closed compact subset  $A \subset \mathbb{R}^d$ , we use  $\text{vol}(A)$  to denote the volume of  $A$ .

For any  $r \geq 0$ ,  $B(x, r)$  denotes the  $d$ -dimensional ball centered at  $x$  with radius  $r$ . For any  $0 \leq n \leq d$ , we use  $B_r^n$  to denote a  $n$ -dimensional ball with radius  $r$  centered at the origin. Define the function  $I(n) =$

\*Research supported by Research Grant Council, Hong Kong, China (project no. 612005).

<sup>†</sup>Department of Computer Science and Engineering, HKUST, Clear Water Bay, Hong Kong. Email: {scheng, chiunk}@cse.ust.hk

$\int_0^\pi (\sin \theta)^n d\theta$  for any integer  $n \geq 0$ . Define the recursive function  $\alpha(n) = I(n)\alpha(n-1)$  for any integer  $n \geq 2$  and  $\alpha(1) = 2$ . Then  $\text{vol}(\mathbb{B}_r^n) = r^n \alpha(n)$ .

For any  $0 \leq n \leq d$ , a  $n$ -flat is a  $n$ -dimensional subspace of  $\mathbb{R}^d$  congruent to  $\mathbb{R}^n$ . Given any subset  $A$  of  $\mathbb{R}^d$ , we use  $\text{aff}(A)$  to denote the flat of the lowest dimension containing  $A$ . We use  $H^n$  to denote a  $n$ -dimensional linear subspace, i.e.,  $H^n$  is a  $n$ -flat passing through the origin. Given  $H^n$  and some  $r > 0$ , define  $H_r^n = \{x \in \mathbb{R}^d : \text{distance from } x \text{ to } H^n \text{ is at most } r\}$ .

We use  $M$  to denote a smooth manifold without boundary embedded in  $\mathbb{R}^d$ . A *medial ball* of  $M$  is a  $d$ -dimensional ball  $B$  such that  $B$  touches  $M$  at two or more points and the interior of  $B$  avoids  $M$ . The *medial axis* of  $M$  is the set of medial ball centers. For any point  $x \in M$ , the local feature size  $\text{lfs}(x)$  is the distance from  $x$  to the medial axis of  $M$ . The *reach* of  $M$ , denoted by  $\gamma(M)$ , is the minimum local feature size over all points in  $M$ . We assume that  $\gamma(M) > 0$ . For any point  $p \in M$ , we use  $T_p(M)$  to denote the  $m$ -flat tangent to  $M$  at  $p$ .

The input points in  $P$  are drawn from  $M$  according to a Poisson process with an unknown parameter  $\lambda$ . A finite point set  $Q \subset M$  is an  $\varepsilon$ -sample for some  $\varepsilon \in (0, 1)$  if for any point  $x \in M$ , there is a point  $p \in Q$  such that  $\|p-x\| \leq \varepsilon \gamma(M)$ . The set  $Q$  is an  $(\varepsilon, \delta)$ -sample for some  $0 < \delta < \varepsilon < 1$  if  $Q$  is an  $\varepsilon$ -sample and for any two points  $p, q \in Q$ ,  $\|p-q\| \geq \delta \gamma(M)$ .

Let  $S \subset \mathbb{R}^d$  be a finite point set. A *weight assignment to  $S$*  associates a real non-negative weight  $w_p$  with each point  $p \in S$ . For any point  $x \in \mathbb{R}^d$ , its *weighted distance* from a point  $p \in S$  with weight  $w_p$  is equal to  $\|p-x\|^2 - w_p$ . For any real number  $\omega \in (0, 1)$ , we say that the weighted point set  $S$  has *weight property*  $[\omega]$  if, for any  $p \in S$ ,  $\sqrt{w_p}$  is at most  $\omega$  times the nearest neighbor distance of  $p$ .

Given a  $d$ -simplex  $\tau$  with vertices in  $S$ , there is a unique point  $x \in \mathbb{R}^d$  at the same weighted distance  $X$  to the vertices of  $\tau$ . The point  $x$  is called the *orthocenter* of  $\tau$  and the ball  $B(x, \sqrt{X})$  is called the *orthoball* of  $\tau$ . Notice that if we view a vertex  $p$  of  $\tau$  as the ball  $B(p, \sqrt{w_p})$ , where  $w_p$  is its weight, the two balls  $B(p, \sqrt{w_p})$  and  $B(x, \sqrt{X})$  intersect at right angle, i.e.,  $\|p-x\|^2 = w_p + X$ .

The simplex  $\tau$  is *weighted Delaunay* if the orthocenter of  $\tau$  is at a smaller weighted distance to the vertices of  $\tau$  than other points in  $S$ . The collection of weighted Delaunay  $d$ -simplices and their boundary simplices form a *weighted Delaunay triangulation* of  $S$ . The *restricted weighted Delaunay triangulation* of  $S$  with respect to a manifold  $M$  is the subset of weighted Delaunay simplices whose dual weighted Voronoi cells intersect  $M$ .

### 3 Slivers.

Cheng et al. [4] showed that, by avoiding *slivers* in the restricted weighted Delaunay triangulation of an  $(\varepsilon, \delta)$ -sample  $S$  of  $M$  where  $\varepsilon/\delta = O(1)$ , one can obtain a triangulation of  $S$  homeomorphic to  $M$ . For any  $\sigma \in (0, 1)$ , we define:

- (i) Vertices and edges are not  $\sigma$ -slivers.
- (ii) For  $2 \leq j \leq d$ , a  $j$ -simplex  $\tau$  is a  $\sigma$ -sliver if  $\text{vol}(\tau) \leq \sigma^j L^j / j!$ , where  $L$  is its longest edge length.<sup>1</sup>

The result in [4] gives us some useful geometric properties and allows us to avoid  $\sigma$ -slivers for any  $\sigma < \sigma_0(\delta/\varepsilon)$ , where  $\sigma_0$  is a function of the ratio  $\delta/\varepsilon$ . The precise statement is given below.

**THEOREM 3.1.** ([4]) *Let  $M \subset \mathbb{R}^d$  be a manifold of dimension  $m$ . Let  $S$  be an  $(\varepsilon, \delta)$ -sample of  $M$  for some  $0 < \delta < \varepsilon < 1$  such that  $\delta = \Omega(\varepsilon)$ . For any  $0 < \sigma < \sigma_0(\delta/\varepsilon)$ , there exists  $\varepsilon_0 = O(\sigma^{O(m^2)})$  such that if  $\varepsilon < \varepsilon_0$ , there is a weight assignment to  $S$  with weight property  $[\omega]$  for some  $\omega < 1/2$  such that:*

- (i) *The restricted weighted Delaunay triangulation of  $S$  is homeomorphic to  $M$ .*
- (ii) *For  $1 \leq j \leq m$ , no  $j$ -simplex in the restricted weighted Delaunay triangulation is a  $\sigma$ -sliver.*
- (iii) *Let  $p$  be a point in  $S$ . Let  $\tau$  be a simplex inside  $B(p, \varepsilon \gamma(M))$  with  $p$  and some other points in  $S$  as vertices. If  $\dim(\tau) = m+1$  and no boundary simplex of  $\tau$  is a  $\sigma$ -sliver,  $\tau$  is a  $\sigma$ -sliver.*

The result in [4] is formulated for the *locally adaptive* case in which  $(\varepsilon, \delta)$ -sampling means that: (i) for any point  $x \in M$ , there is a point  $p \in S$  such that  $\|p-x\| \leq \varepsilon \text{lfs}(x)$ ; (ii) for any points  $p, q \in S$ ,  $\|p-q\| \geq \delta \text{lfs}(p)$ . The notion of  $(\varepsilon, \delta)$ -sampling used in this paper is non-adaptive in the sense that  $\text{lfs}(x)$  in (i) and  $\text{lfs}(p)$  in (ii) above are substituted by the reach  $\gamma(M)$ . It is standard in the literature that results sensitive to the local feature sizes can be carried over to become results sensitive to the reach.

A bound of  $O((\delta/\varepsilon)^{O(d^2)})$  is stated for  $\sigma_0(\delta/\varepsilon)$  in [4]. This can be easily improved to  $O((\delta/\varepsilon)^{O(m^2)})$  and we give the arguments in the appendix.

Theorem 3.1 does not subsume our work because our sample set  $P$  needs not be an  $(\varepsilon, \delta)$ -sample and one cannot compute the restricted weighted Delaunay triangulation without knowing  $M$ .

<sup>1</sup>In [2, 4], the shortest edge length of  $\tau$  is used. The difference is minor as the longest and shortest edge lengths are known to be within a constant factor of each other for an  $(\varepsilon, \delta)$ -sample where  $\varepsilon/\delta = O(1)$ . Using the longest edge length in the experiments alleviates the issue of encountering a tiny shortest edge length.

#### 4 Basic algorithm.

Our basic algorithm picks some trial points and collect point samples in their neighborhoods. For  $j \geq 2$  and for each trial point  $p$ , we check if  $p$  and other points in the neighborhood can form a  $j$ -dimensional non-sliver simplex such that its boundary simplices are also not slivers. By Theorem 3.1, this is possible only if  $j \leq m$ . This is the intuition why simplicial shape analysis can detect the dimension of  $M$ . The details of the basic algorithm are presented in the following.

For any  $j, n \geq 0$ , let  $f(j, n)$  and  $g(n)$  be two functions whose values are between  $\ln n$  and  $n^{1/4}$ . We will show how to choose  $f(j, n)$  and  $g(n)$  later. For any  $1 < n \leq d, r > 0$  and  $c \in (0, 1)$ , define

$$\text{Ratio}(n, r, c) = \max \left\{ \begin{array}{l} \text{vol}(\mathbf{B}_r^n \cap \mathbf{H}_{cr}^{n-1}) / \text{vol}(\mathbf{B}_r^n) \\ \text{vol}(\mathbf{B}_r^{n+1} \cap \mathbf{H}_{cr}^{n-1}) / \text{vol}(\mathbf{B}_r^{n+1}) \end{array} \right.$$

assuming that  $\mathbf{H}^{n-1} \subset \text{aff}(\mathbf{B}_r^n) \subset \text{aff}(\mathbf{B}_r^{n+1})$ . If  $n = d$ ,  $\text{vol}(\mathbf{B}_r^{n+1} \cap \mathbf{H}_{cr}^{n-1}) / \text{vol}(\mathbf{B}_r^{n+1})$  is omitted from the definition. Let  $\sigma_1(j)$  be a function whose value is in  $(0, 1/2)$ . Let  $k$  be some fixed positive integer. We show later how to choose  $\sigma_1(j)$  and  $k$ .

Our basic algorithm DIMENSION calls ESTIMATE( $P$ )  $k$  times. Each call returns a dimension estimate which is stored in an array  $A$ . The most frequent value in  $A$  is the answer.

DIMENSION( $P, k$ )

1. For  $i := 1$  to  $k$ ,  $A[i] := \text{ESTIMATE}(P)$ .
2. Return the value with the highest frequency in  $A$ .

ESTIMATE( $P$ )

1. Initialize  $j = 2$ .
2. Draw a set  $K$  of  $g(|P|)$  trial points from  $P$  uniformly at random.
3. If TRIAL( $K, j$ ) returns FAIL, return  $j - 1$ . Otherwise, increment  $j$  and go to step 2.

TRIAL( $K, j$ )

1. For each trial point  $p \in K$ ,
  - (a) Compute the work zone  $Z_p$  consisting of  $p$  and the  $f(j, |P|)$  nearest neighbors of  $p$  in  $P$ . Let  $r_p$  be the distance between  $p$  and the furthest point in  $Z_p$ . Draw  $q$  uniformly at random among the  $f(j, |P|) - 1$  nearest neighbors of  $p$ .
  - (b) Find a  $(j - 1)$ -simplex  $\tau_{j-1}$  such that: (i)  $\tau_{j-1}$  has  $p$  and some other points in  $Z_p \setminus \{q\}$  as vertices, (ii) its shortest edge length is at least  $r_p/20$ ; (iii)  $\tau_{j-1}$  and its boundary simplices are not  $\sigma_1(j)$ -slivers.

- (c) If  $\tau_{j-1}$  is found and  $\text{vol}(\tau_{j-1} * q) \leq \sigma_1(j)^j L^j / j!$ , where  $L$  is the longest edge length of  $\tau_{j-1}$ , call  $\tau_{j-1} * q$  a bad simplex and set  $N_{p,j} = 1$ . If  $\tau_{j-1}$  is not found or  $\tau_{j-1} * q$  is not bad, set  $N_{p,j} = 0$ .

2. Return FAIL if the sum  $\sum_{p \in K} N_{p,j}$  is greater than  $(2e + 1)e^{3/5} \cdot \sum_{p \in K} \text{Ratio}(j, r_p, 2\sigma_1(j))$ . Otherwise, return PASS.

DIMENSION and ESTIMATE together make repeated calls to TRIAL. Each repetition is an independent trial and so the repetitions increase the success probability. TRIAL finds a non-sliver  $\tau_{j-1}$  and then tests if  $\tau_{j-1} * q$  is a bad simplex for a point  $q$  drawn uniformly at random. Under the condition that  $j \leq m$ , we will prove that the expected number of bad simplices obtained is bounded by  $e^{3/5} \sum_{p \in K} \text{Ratio}(j, r_p, 2\sigma_1(j))$ . By the Chernoff bound, it is unlikely in step 2 of TRIAL for the actual number of bad simplices to exceed  $(2e + 1)e^{3/5} \sum_{p \in K} \text{Ratio}(j, r_p, 2\sigma_1(j))$ . Conversely, if this happens, it is likely that  $j$  has exceeded the manifold dimension, i.e.,  $j \geq m + 1$ .

The above intuition will be captured in a formal analysis in the next section. We analyze the running time below.

LEMMA 4.1. *The running time of the function call TRIAL( $K, j$ ) is  $O(d|P|g(|P|) + dj^2 f(j, |P|)^{j+1} g(|P|) + j^3 f(j, |P|)^{j+1} g(|P|) \log f(j, |P|))$ .*

*Proof.* In step 1(a), the work zone of a trial point  $p$  can be computed in  $O(d|P|)$  time using a linear-time selection algorithm. We discuss the implementation of step 1(b).

We assign vertex ids from the range  $[0, f(j, |P|) - 1]$  to the points in  $Z_p \setminus \{q\}$ . A simplex  $\tau$  with vertices in  $Z_p \setminus \{q\}$  is represented by the increasing sequence of its vertex ids  $v_0 < v_1 < \dots < v_j$ . We store  $\tau$  in a dictionary using this vertex id sequence as the key. Notice that comparing the keys of two simplices takes  $O(j)$  time.

We enumerate simplices with increasing dimension using  $Z_p \setminus \{q\}$  as the vertex set. Assume that we have inductively enumerated all  $i$ -simplices for some  $1 \leq i \leq j - 1$  and mark the  $i$ -simplices such that they and their boundary simplices are not  $\sigma_1(j)$ -slivers. We store these simplices in a dictionary  $D$  indexed by their keys. The size of  $D$  is  $O(f(j, |P|)^{i+1})$ . Next, we enumerate simplices of dimension  $i + 1$ . There are  $\binom{f(j, |P|)}{i+2} < f(j, |P|)^{i+2}$  of them and they can be enumerated in  $O(f(j, |P|)^{i+2})$  time. These  $(i + 1)$ -simplices can be stored in a dictionary  $D'$  indexed by their keys. For each  $(i + 1)$ -simplex  $\tau$ , we compute its volume using

the QR decomposition and the Householder transformation [12], which takes  $O(d(i+1)^2)$  time for simplices with dimension  $i + 1$ . Then, we look up the  $i + 2$  boundary  $i$ -simplices of  $\tau$  in the dictionary  $D$  in  $O(i^3 \log f(j, |P|))$  time. If  $\tau$  is not a  $\sigma_1(j)$ -sliver and the boundary  $i$ -simplices of  $\tau$  are marked in  $D$ , we mark the entry of  $\tau$  in  $D'$ . After processing all entries of  $D'$ , we overwrite  $D$  by  $D'$  and repeat the above until  $i = j$ .

Enumerating simplices takes  $O(\sum_{i=1}^j \binom{f(j, |P|)}{i+1} di^2 + \sum_{i=1}^j \binom{f(j, |P|)}{i+1} i^3 \log f(j, |P|)) = O(\sum_{i=1}^j di^2 f(j, |P|)^{i+1} + \sum_{i=1}^j i^3 f(j, |P|)^{i+1} \log f(j, |P|))$  time. So step 1(b) takes  $O(dj^2 f(j, |P|)^{j+1} + j^3 f(j, |P|)^{j+1} \log f(j, |P|))$  time. The volume computation in Step 1(c) takes  $O(dj^2)$  time. In all, step 1 takes  $O(dj^2 f(j, |P|)^{j+1} g(|P|) + j^3 f(j, |P|)^{j+1} g(|P|) \log f(j, |P|))$  for all trial points.

One can derive a formula for  $\text{vol}(\mathbf{B}_r^n \cap \mathbf{H}_{cr}^{n-1})$  that can be evaluated in  $O(n)$  time using integration by parts. The volume of  $\mathbf{B}_r^n$  can also be evaluated in  $O(n)$  time using integration by parts. So  $\text{Ratio}(j, r_p, 2\sigma_1(j))$  can be computed in  $O(j)$  time for each  $p \in K$ . Thus, step 2 takes  $O(jg(|P|))$  time for all trial points.  $\square$

## 5 Analysis.

The analysis of our algorithm consists of several steps given in the following subsections.

**5.1 Technical results.** We introduce several technical results. The first one follows from the Chernoff bounds [17]. Lemma 5.2 and 5.4 follow from standard derivation. Lemma 5.3 follows from a result in [11].

**LEMMA 5.1.** *Let  $X$  be the sum of some independent random binary variables  $X_1, X_2, \dots$ . Assume that  $E[X] = \mu$ . Then, (i)  $\Pr(X < (1-t)\mu) < \exp\left(-\frac{t^2\mu}{2}\right)$  for  $t \in (0, 1]$ ; (ii) if  $\Delta \geq \mu$ , then  $\Pr(X > (1+t)\Delta) < 2^{-t\Delta}$  for  $t > 2e - 1$ .*

**LEMMA 5.2.** *For any  $1 < n < m$ ,  $r > 0$  and  $c \in (0, 1)$ , if  $\mathbf{H}^{n-1} \subset \text{aff}(\mathbf{B}_r^n) \subset \text{aff}(\mathbf{B}_r^{n+1}) \subseteq \text{aff}(\mathbf{B}_r^m)$ , then  $\text{Ratio}(n, r, c) > \text{vol}(\mathbf{B}_r^m \cap \mathbf{H}_{cr}^{n-1}) / \text{vol}(\mathbf{B}_r^m)$ .*

**LEMMA 5.3.** *Let  $p$  be a point in  $\mathbf{M}$ . Let  $r = t\gamma(\mathbf{M})$  for some  $t \in (0, 1)$ . Then,  $\text{vol}(\mathbf{M} \cap B(p, r)) \geq \text{vol}(\mathbf{B}_\rho^m)$ , where  $\rho = (\sqrt{1 - t^2/4})r$ .*

**LEMMA 5.4.** *Let  $p$  be a point in  $\mathbf{M}$ . Let  $t \in (0, 1/(2 + 8\sqrt{m})]$  be a real number. Let  $dV$  denote the differential volume element of  $\mathbf{M}$  at a point  $y \in \mathbf{M} \cap B(p, t\gamma(\mathbf{M}))$ . Assume that  $\mathbf{T}_p(\mathbf{M}) = \{(x_1, \dots, x_m, 0, \dots, 0) : -\infty < x_i < \infty\}$ . Then,  $dV \leq \sqrt{e} \cdot dx_1 \dots dx_m$ .*

**5.2 Work zones.** We prove that the radius of each work zone is less than  $\varepsilon_1 \gamma(\mathbf{M})$  where  $\varepsilon_1 = O((f(j, |P|)/|P|)^{1/m})$ . Because the work zone is small, the portion of  $\mathbf{M}$  inside a work zone is rather flat. Thus, if a simplex connecting point samples in the work zone has dimension greater than  $m$ , it is intuitive that the simplex is a sliver. An analogous situation is a tetrahedron connecting four points on a plane.

**LEMMA 5.5.** *Assume that  $|P|$  is large enough that  $|P| \text{vol}(\mathbf{B}_r^m) = 4(f(j, |P|) + 1) \text{vol}(\mathbf{M})$  for some  $r = (\sqrt{1 - \varepsilon_1^2/4})\varepsilon_1 \gamma(\mathbf{M})$  and  $\varepsilon_1 \in (0, 1)$ . For any point  $p \in \mathbf{M}$ ,  $\Pr(|P \cap \mathbf{M} \cap B(p, \varepsilon_1 \gamma(\mathbf{M}))| \geq f(j, |P|) + 1) > 1 - (e|P|)^{-1}$ .*

*Proof.* Let  $P = \{p_1, p_2, \dots\}$ . For any  $1 \leq i \leq |P|$ , define a random variable  $X_i = 1$  if  $p_i$  belongs to  $B(p, \varepsilon_1 \gamma(\mathbf{M}))$  and  $X_i = 0$  otherwise. Clearly,  $\Pr(X_i) = \text{vol}(\mathbf{M} \cap B(p, \varepsilon_1 \gamma(\mathbf{M}))) / \text{vol}(\mathbf{M})$ . Define  $X = \sum_{i=1}^{|P|} X_i$ . Our assumption and Lemma 5.3 imply that  $E[X] = |P| \text{vol}(\mathbf{M} \cap B(p, \varepsilon_1 \gamma(\mathbf{M}))) / \text{vol}(\mathbf{M}) \geq |P| \text{vol}(\mathbf{B}_r^m) / \text{vol}(\mathbf{M}) = 4(f(j, |P|) + 1)$ . By Lemma 5.1(i),  $\Pr(X < f(j, |P|) + 1) < \Pr(X < (1 - 1/\sqrt{2})E[X]) < \exp(-f(j, |P|) - 1) \leq (e|P|)^{-1}$  as  $f(j, |P|) \geq \ln |P|$ .  $\square$

By Lemma 5.5, the radius of a work zone is no more than  $\varepsilon_1 \gamma(\mathbf{M})$  with high probability. The small radii of the work zones imply that they are disjoint with high probability as stated in the result below.

**LEMMA 5.6.** *Assume that  $|P|$  is large enough that  $|P| \text{vol}(\mathbf{B}_r^m) = 4(f(j, |P|) + 1) \text{vol}(\mathbf{M})$  for some  $r = (\sqrt{1 - \varepsilon_1^2/4})\varepsilon_1 \gamma(\mathbf{M})$  and  $\varepsilon_1 < 1/(4 + 16\sqrt{m})$ . Among the work zones of trial points used in  $\text{TRIAL}(K, j)$  for  $2 \leq j \leq 2m$  over all  $k$  iterations in  $\text{DIMENSION}$ , it holds with probability  $1 - O(k^2 m^2 2^{4m} |P|^{-1/4})$  that no two work zones overlap.*

**5.3 Bad simplex.** Next, we prove a bound on the probability of getting a bad simplex  $\tau_{j-1} * q$  in step 1(c) of  $\text{TRIAL}$ , provided that  $j \leq m$ . This probability bound leads to the bound  $e^{3/5} \sum_{p \in K} \text{Ratio}(j, r_p, 2\sigma_1(j))$  on the expected number of bad simplices in step 2 of  $\text{TRIAL}$ . Thus, it is unlikely for the number of bad simplices to exceed  $(2e + 1)e^{3/5} \sum_{p \in K} \text{Ratio}(j, r_p, 2\sigma_1(j))$  by the Chernoff bound. If this happens, we can conclude that  $j$  has exceeded the manifold dimension, i.e.,  $j \geq m + 1$ .

The volume of  $\tau_{j-1} * q$  is determined by the distance between  $q$  and  $\text{aff}(\tau_{j-1})$ . We first bound the probability of  $q$  being close to  $\text{aff}(\tau_{j-1})$  in the next lemma.

**LEMMA 5.7.** *Let  $t \in (0, 1/(2 + 8\sqrt{m})]$  be a real number. Let  $\tau_{j-1}$  be a  $(j - 1)$ -simplex with  $p$  and*

some other points in  $M \cap B(p, r)$  as vertices, where  $r = t\gamma(M)$ . If  $x$  is a point drawn at random from  $M \cap B(p, r)$ , then for any  $c \in (0, 1)$  and any  $H^{j-1} \subset \text{aff}(B_r^m)$ ,  $\Pr(\text{dist}(x, \text{aff}(\tau_{j-1})) \leq cr \mid j \leq m) < e^{3/5} \text{vol}(B_r^m \cap H_{cr}^{j-1}) / \text{vol}(B_r^m)$ .

*Proof.* Translate space so that  $p$  is at the origin. Define  $Y = \{y \in M \cap B(p, r) : \text{dist}(y, \text{aff}(\tau_{j-1})) \leq cr\}$ . Let  $Y'$  denote the projection of  $Y$  onto  $T_p(M)$  and for any point  $y \in Y$ , let  $y' \in Y'$  denote the projection of  $y$ . For any point  $y' \in Y'$ ,  $\|p - y'\| \leq r$ . The projection of  $\text{aff}(\tau_{j-1})$  onto  $T_p(M)$  is a linear subspace of dimension  $j - 1$  or less. Since projection either preserves or decreases the distance between two points, the distance between  $y'$  in  $Y'$  and the projection of  $\text{aff}(\tau_{j-1})$  is at most  $cr$ . Hence, we can find a  $(j - 1)$ -dimensional linear subspace  $H^{j-1} \subset T_p(M)$  such that  $y' \in H_{cr}^{j-1}$  for any point  $y' \in Y'$ . Let  $B_r^m$  be a  $m$ -ball in  $T_p(M)$  such that  $H^{j-1} \subset \text{aff}(B_r^m)$ . So  $Y' \subseteq B_r^m \cap H_{cr}^{j-1}$ . It suffices to prove the probability bound for these choices of  $H^{j-1}$  and  $B_r^m$  because the probability bound is insensitive to the choices as long as  $H^{j-1} \subset \text{aff}(B_r^m)$ . Let  $\rho = (\sqrt{1 - t^2/4})r$ . By Lemma 5.3,  $\text{vol}(M \cap B(p, r)) \geq \text{vol}(B_\rho^m)$ . By Lemma 5.4,  $\text{vol}(Y) \leq \sqrt{e} \text{vol}(Y')$ . Combining these results, we conclude that  $\Pr(\text{dist}(x, \text{aff}(\tau_{j-1})) \leq cr \mid j \leq m) = \text{vol}(Y) / \text{vol}(M \cap B(p, r)) \leq \sqrt{e} \text{vol}(Y') / \text{vol}(B_\rho^m) \leq \sqrt{e} \cdot (\text{vol}(B_r^m \cap H_{cr}^{j-1}) / \text{vol}(B_r^m)) \cdot (\text{vol}(B_r^m) / \text{vol}(B_\rho^m))$ . Using the fact that  $\rho = (\sqrt{1 - t^2/4})r$ , the ratio  $\text{vol}(B_r^m) / \text{vol}(B_\rho^m)$  becomes  $(1 - t^2/4)^{-m/2}$ . Then, further simplification gives  $\Pr(\text{dist}(x, \text{aff}(\tau_{j-1})) \leq cr \mid j \leq m) < e^{3/5} \text{vol}(B_r^m \cap H_{cr}^{j-1}) / \text{vol}(B_r^m)$ .  $\square$

Next, we use Lemma 5.2 and Lemma 5.7 to bound the probability of getting a bad simplex.

**LEMMA 5.8.** *Let  $t \in (0, 1/(2 + 8\sqrt{m}))$  be a real number. Let  $\tau_{j-1}$  be a  $(j - 1)$ -simplex with  $p$  and some other points in  $M \cap B(p, r)$  as vertices, where  $r = t\gamma(M)$ . Let  $L$  be the longest edge length of  $\tau_{j-1}$ . Assume that  $\text{vol}(\tau_{j-1}) > \sigma^{j-1} L^{j-1} / (j - 1)!$  and  $x$  is a point drawn at random from  $M \cap B(p, r)$ . Then, for any  $c \in (0, 1)$  and for any  $\sigma \in (0, c/2]$ ,  $\Pr(\text{vol}(\tau_{j-1} * x) \leq \frac{\sigma^j L^j}{j!} \mid j \leq m) < e^{3/5} \text{Ratio}(j, r, c)$ .*

*Proof.* Since  $L \leq 2r$  and  $\text{vol}(\tau_{j-1}) > \sigma^{j-1} L^{j-1} / (j - 1)!$ , we get  $\frac{\sigma^j L^j}{(j-1)! \cdot \text{vol}(\tau_{j-1})} \leq 2\sigma r \leq cr$ . Note that  $\text{dist}(x, \text{aff}(\tau_{j-1})) = j \cdot \text{vol}(\tau_{j-1} * x) / \text{vol}(\tau_{j-1})$ . So  $\text{vol}(\tau_{j-1} * x) \leq \sigma^j L^j / j!$  implies  $\text{dist}(x, \text{aff}(\tau_{j-1})) \leq cr$ .

What is the probability that  $\text{dist}(x, \text{aff}(\tau_{j-1})) \leq cr$ ? Because  $\text{Ratio}(j, r, c) \geq \text{vol}(B_r^j \cap H_{cr}^{j-1}) / \text{vol}(B_r^j)$ , if  $j = m$ , we can apply Lemma 5.7 and conclude that  $\Pr(\text{dist}(x, \text{aff}(\tau_{j-1})) \leq cr \mid j \leq m) <$

$e^{3/5} \text{vol}(B_r^m \cap H_{cr}^{m-1}) / \text{vol}(B_r^m) \leq e^{3/5} \text{Ratio}(m, r, c)$ . If  $j < m$ , Lemma 5.2 and Lemma 5.7 imply that  $e^{3/5} \text{Ratio}(j, r, c) > e^{3/5} \text{vol}(B_r^m \cap H_{cr}^{j-1}) / \text{vol}(B_r^m) > \Pr(\text{dist}(x, \text{aff}(\tau_{j-1})) \leq cr \mid j \leq m)$ .  $\square$

**5.4 Sparse sample.** We have so far established an upper bound on the expected number of bad simplices in step 2 of TRIAL under the condition that  $j \leq m$ . This will allow us to show that the probability of underestimation is small. We need to show that the probability of overestimating the manifold dimension is small too. The idea is to invoke Theorem 3.1(iii) and show that, when  $j \geq m + 1$ , the number of bad simplices is likely to exceed the bound  $(2e + 1)e^{3/5} \sum_{p \in K} \text{Ratio}(j, r_p, 2\sigma_1(j))$  in step 2 of TRIAL.

Because Theorem 3.1 is applicable to a sparse set of point samples, we first show that a sparse sample of  $M$  with nice properties can be found for each trial point. Recall that for a trial point  $p$ ,  $r_p$  is the distance between  $p$  and the furthest point in  $Z_p$ .

**LEMMA 5.9.** *Let  $\beta_1 = 80^{-m} e^{-1/2} / 32$ . There is a value  $n_0 = 2^{\Theta(m^6)}$  such that when  $|P| > n_0$ , for any trial point  $p$  and any subset  $V \subseteq Z_p$  such that the interpoint distance in  $V$  is at least  $r_p/20$ , there exists a point set  $Q_V \subset M$  with probability  $1 - O(80^m |P|^{-1} + 80^m e^{-\beta_1 f(j, |P|)})$  such that  $V \subset Q_V$ ,  $Q_V \cap B(p, r_p/4) = P \cap B(p, r_p/4)$ , and  $Q_V$  is an  $(\varepsilon/16, \varepsilon/40)$ -sample of  $M$ , where  $\varepsilon\gamma(M) = r_p$ .*

*Proof.* By our assumption,  $|P|$  is large enough that  $|P| \frac{\text{vol}(B_r^m)}{\text{vol}(M)} = 4(f(j, |P|) + 1) \text{vol}(M)$  for some  $r = (\sqrt{1 - \varepsilon_1^2/4})\varepsilon_1\gamma(M)$  and  $\varepsilon_1 \in (0, 1)$ . Let  $r_0 = \varepsilon_1\gamma(M)$ . By Lemma 5.5,  $r_p = \varepsilon\gamma(M)$  for some  $\varepsilon \leq \varepsilon_1$  with probability at least  $1 - (e|P|)^{-1}$ . Let  $r_1 = r_p/40$ . Let  $r_2 = r_1/2$ . Denote by  $S$  the  $f(j, |P|) - 2$  points in  $Z_p$  other than  $p$ , the furthest point in  $Z_p$  from  $p$ , and the point  $q$  drawn from  $Z_p$  in step 1(a) of TRIAL. The points in  $S$  are uniformly distributed in  $M \cap B(p, r_p)$ .

Let  $\rho = (\sqrt{1 - \varepsilon_1^2/25600})r_2$ . For any point  $x \in M \cap B(p, r_p/2)$ , Lemma 5.3 and Lemma 5.4 imply that

$$\frac{\text{vol}(M \cap B(x, r_2))}{\text{vol}(M \cap B(p, r_p))} \geq \frac{\text{vol}(B_\rho^m)}{\sqrt{e} \text{vol}(B_{r_p}^m)} > \frac{1}{2} 80^{-m} e^{-1/2}.$$

Lemma 5.4 requires  $\varepsilon_1 = O(m^{-1/2})$  and hence  $|P| = \Omega(m^{m/2})$ . This holds because  $|P| > n_0 = 2^{\Theta(m^6)}$ . We have  $E[|S \cap B(x, r_2)|] > \frac{1}{2} 80^{-m} e^{-1/2} (f(j, |P|) - 2) > \frac{1}{4} 80^{-m} e^{-1/2} f(j, |P|)$ . By Lemma 5.1(i),  $\Pr(|S \cap B(x, r_2)| \leq \frac{1}{2} E[|S \cap B(x, r_2)|]) = O(e^{-\beta_1 f(j, |P|)})$ . It follows that for any point  $x \in B(p, r_p/2)$ ,  $B(x, r_2)$  contains a point in  $S$  with probability at least  $(1 -$

$$e^{-1}|P|^{-1}) \cdot (1 - O(e^{-\beta_1 f(j, |P|)})) = 1 - O(|P|^{-1} + e^{-\beta_1 f(j, |P|)}).$$

We take a maximal packing of balls  $B(x, r_1)$  with centers in  $M$  such that, for every point  $v \in V$ ,  $B(v, r_1)$  is one of the balls in the packing. Note that for any  $u, v \in V$ ,  $B(u, r_1)$  and  $B(v, r_1)$  are disjoint because  $\|u - v\| \geq r_p/20$  by assumption. For each ball  $B(x, r_1)$  in the packing, if there is a point in  $P$  inside  $B(x, r_1/2) = B(x, r_2)$ , we pick one and put it in a set  $Q$ . For every  $v \in V$ ,  $v$  is the point to be picked inside  $B(v, r_1)$ . For each ball  $B(x, r_1)$  in the packing such that  $B(x, r_2)$  does not contain any point in  $P$ , we just pick an arbitrary point in  $M \cap B(x, r_2)$  and put it in  $Q$ . It is clear that  $V \subset Q_V$  by construction.

Since  $\text{vol}(M \cap B(x, r_2)) > \frac{1}{2} 80^{-m} e^{-1/2} \text{vol}(M \cap B(p, r_p))$ , the number of balls in the packing with centers inside  $B(p, r_p/2) < 2\sqrt{e} 80^m$ . So it holds with probability  $1 - O(80^m |P|^{-1} + 80^m e^{-\beta_1 f(j, |P|)})$  that every ball  $B(x, r_2)$  in the packing, where  $x \in B(p, r_p/2)$ , contains a point in  $S$ . That is,  $Q_V$  contains exactly one point in  $S$  in every ball  $B(x, r_1)$  in the packing such that  $x \in B(p, r_p/2)$ . Because  $r_1 = r_p/40$ , each point in  $Q_V \cap B(p, r_p/4)$  must belong to some ball  $B(x, r_1)$  such that  $x \in B(p, r_p/2)$ . It follows that  $Q_V \cap B(p, r_p/4) = P \cap B(p, r_p/4)$ .

Two points in  $Q_V$  are at distance  $r_1 = (\varepsilon/40) \gamma(M)$  or more away by construction. Due to the maximality of the ball packing, any point  $z$  in  $M$  must lie inside  $B(x, 2r_1)$  for some ball  $B(x, r_1)$  in the packing. So  $Q_V$  has a point in  $B(x, r_2)$  that is within a distance of  $5r_2$  from  $z$ . Since  $5r_2 = (\varepsilon/16) \gamma(M)$ ,  $Q_V$  is an  $(\varepsilon/16)$ -sample. Hence,  $Q_V$  is an  $(\varepsilon/16, \varepsilon/40)$ -sample.  $\square$

To conclude that a simplex  $\tau$  tested in step 1(c) of TRIAL is a bad simplex by Theorem 3.1(iii), the boundary simplices of  $\tau$  must be non-slivers. This can be ensured by Theorem 3.1(ii) if  $\tau$  belongs to the restricted weighted Delaunay triangulation. Because TRIAL only examines simplices with vertices in  $Z_p$ , it is necessary that  $Z_p$  contains the vertices of any simplex incident to  $p$  in the restricted weighted Delaunay triangulation. The next lemma proves this result.

**LEMMA 5.10.** *Let  $Q_p$  denote  $Q_V$  in Lemma 5.9 for the case of  $V$  containing the trial point  $p$  only. For any  $\omega < 1/2$  and for any weight assignment to  $Q_p$  with weight property  $[\omega]$ , simplices incident to  $p$  in the restricted weighted Delaunay triangulation of  $Q_p$  have vertices in  $Z_p$ .*

*Proof.* Take some weight assignment of  $Q_p$  with weight property  $[\omega]$ . Take a restricted weighted Delaunay simplex  $\tau$  incident to  $p$ . Its dual weighted Voronoi cell intersects  $M$  at some point  $z$ . The orthoball of  $\tau$

centered at  $z$  is empty and so its radius is less than  $(\varepsilon/16) \gamma(M)$ . Take any point  $q$  in  $Q_p$ . The bisector plane between  $q$  and its nearest neighbor  $s$  in  $Q_p$  intersects  $M$  at some point  $y$ . Moreover,  $q$  and  $s$  cannot be further away from  $y$  than  $(\varepsilon/16) \gamma(M)$ . It follows that the nearest neighbor distance of  $q$  is at most  $(\varepsilon/8) \gamma(M)$ . So each weighted vertex of  $\tau$  can be viewed as a ball with radius less than  $(\varepsilon/16) \gamma(M)$ . This ball must intersect the orthoball of  $\tau$  centered at  $z$ . This implies that the distance between  $p$  and any vertex of  $\tau$  is less than  $(\varepsilon/4) \gamma(M) = r_p/4$ . Because  $Q_p \cap B(p, r_p/4) = P \cap B(p, r_p/4)$  by Lemma 5.9, we conclude that the vertices of  $\tau$  belong to  $Z_p$ .  $\square$

More properties of  $Q_p$  follow from Theorem 3.1.

**LEMMA 5.11.** *Let  $\sigma'$  be a value less than  $\sigma_0(1/40)$ . For any  $0 < \sigma < \sigma'$ , there is a value  $n_0 = \sigma^{-\Theta(m^3)}$  such that for any trial point  $p$ , the following hold when  $|P| > n_0$ .*

- (i) *Let  $Q_p$  be the  $(\varepsilon/16, \varepsilon/40)$ -sample of  $M$  in Lemma 5.10 for  $p$ . There is a weight assignment to  $Q_p$  with weight property  $[\omega]$  for some  $\omega < 1/2$  such that the restricted weighted Delaunay triangulation of  $Q_p$  is homeomorphic to  $M$  and for  $1 \leq j \leq m$ , no  $j$ -simplex in this restricted weighted Delaunay triangulation is a  $\sigma$ -sliver.*
- (ii) *Let  $\tau$  be a  $(m + 1)$ -simplex with  $p$  and some other points in  $Z_p$  as vertices such that the shortest edge length of  $\tau$  is at least  $(\varepsilon/20) \gamma(M)$ . If the boundary simplices of  $\tau$  are not  $\sigma$ -slivers,  $\tau$  is a  $\sigma$ -sliver.*

*Proof.* Because  $|P| > n_0 = \sigma^{-\Theta(m^3)}$ , we have  $|P| \text{vol}(B_r^m) = 4(f(j, |P|) + 1) \text{vol}(M)$  for some  $r = (\sqrt{1 - \varepsilon_1^2/4}) \varepsilon_1 \gamma(M)$  and  $\varepsilon_1 = O(\sigma^{-\Theta(m^3)})$ . Since  $\varepsilon \leq \varepsilon_1$ , we can assume that  $n_0$  is large enough that the sampling density condition in Theorem 3.1 is met.

Since  $Q_p$  is an  $(\varepsilon/16, \varepsilon/40)$ -sample, it is also an  $(\varepsilon, \varepsilon/40)$ -sample. So  $\sigma_0(1/40)$  is the threshold of the sliver measure in Theorem 3.1 for  $Q_p$ . Thus, (i) follows from Theorem 3.1(i) and (ii). Note that Theorem 3.1(iii) dictates that we use the sampling density  $\varepsilon = r_p/\gamma(M)$  instead of  $\varepsilon/16$ .

Consider (ii). Let  $V$  be the set of vertices of  $\tau$ . Let  $Q_V$  be the sample guaranteed by Lemma 5.9. Again,  $Q_V$  is an  $(\varepsilon, \varepsilon/40)$ -sample and  $\sigma_0(1/40)$  is the threshold of the sliver measure in Theorem 3.1 for  $Q_V$ . Thus, (ii) follows from Theorem 3.1(iii).  $\square$

**5.5 Theoretical guarantees.** We assemble the previous results to obtain the theoretical guarantees claimed. First, we show that ESTIMATE underestimates the manifold dimension with a small probability.

LEMMA 5.12. Consider the calling of TRIAL( $K, j$ ) for some  $j \in [2, m]$ . Assume that  $\sigma_1(j) < I(j+1)/(4e^{3/5}(2e+1))$ . Let  $\beta_0(j) = 8e^{8/5}\sigma_1(j)(1-4\sigma_1(j)^2)^{(j-1)/2}/(I(2)\log_2 e)$ . There is a value  $n_0 = 2^{\Theta(m^6)}$  such that when  $|P| > n_0$ , ESTIMATE returns  $j-1$  after this call with probability  $O(e^{-\beta_0(j)g(|P|)})$ , assuming disjoint work zones.

*Proof.* Let  $p$  be a trial point in a call TRIAL( $K, j$ ) for some  $j \in [2, m]$ . Let  $q$  be the point drawn from  $Z_p$  at random in step 1(a). Assume that step 1(b) finds a simplex  $\tau_{j-1}$ .

Since  $|P| > n_0$ , we have  $|P| \text{vol}(\mathbf{B}_r^m) = 4(f(j, |P|) + 1) \text{vol}(\mathbf{M})$  for some  $r = (\sqrt{1 - \varepsilon_1^2/4}) \varepsilon_1 \gamma(\mathbf{M})$  and  $\varepsilon_1 = O(m^{-1/2})$ . Thus, we can assume that  $n_0$  is large enough that  $\tau_{j-1}$  lies inside a small neighborhood of  $p$  as required by Lemma 5.8.

By Lemma 5.8,  $\Pr(\text{vol}(\tau_{j-1} * q) < \sigma_1(j)^j L^j / j!) < e^{3/5} \text{Ratio}(j, r_p, 2\sigma_1(j))$ . Assuming disjoint work zones, the expected number of bad simplices obtained in step 2 is less than  $\Delta_j = e^{3/5} \sum_{p \in K} \text{Ratio}(j, r_p, 2\sigma_1(j))$ . A standard derivation allows us to show that  $(2e+1)\Delta_j < g(|P|)$  for our choice of  $\sigma_1(j)$ . By Lemma 5.1(ii), we get  $(2e+1)\Delta_j$  or more bad simplices with probability less than  $2^{-2e\Delta_j}$ . One can also derive that  $2e\Delta_j \geq \beta_0(j)(\log_2 e)g(|P|)$ . So the probability bound is  $O(2^{-2e\Delta_j}) = O(e^{-\beta_0(j)g(|P|)})$ .  $\square$

Next, we prove that ESTIMATE overestimates the manifold dimension with a small probability.

LEMMA 5.13. Let  $\beta_1 = 80^{-m}e^{-1/2}/32$ . Suppose that  $\sigma_1(m+1)$  is less than but of the same order as  $\min(\sigma_0(1/40), I(m+2)/(4e^{3/5}(2e+1)))$ . There is a value  $n_0 = 2^{\Theta(m^6)}$  such that when  $|P| > n_0$ , ESTIMATE overestimates the dimension with probability  $O(80^m g(|P|)|P|^{-1} + 80^m g(|P|)e^{-\beta_1 f(m+1, |P|)})$ , assuming disjoint work zones.

*Proof.* Assume that ESTIMATE calls TRIAL( $K, m+1$ ). Let  $\sigma_2 = (\sigma_1(m+1)/40)^{m+1}$ . By Lemma 5.9 and Lemma 5.11(i), for a trial point  $p$ , it holds with probability  $1 - O(80^m |P|^{-1} + 80^m e^{-\beta_1 f(j, |P|)})$  that step 1(b) finds the simplex  $\tau_m$ . Since  $\sigma_2 < \sigma_1(m+1)$ ,  $\tau_m$  and its boundary simplices are not  $\sigma_2$ -slivers. Since TRIAL uses  $g(|P|)$  trial points, step 1(b) succeeds for every trial point with probability  $1 - O(80^m g(|P|)|P|^{-1} + 80^m g(|P|)e^{-\beta_1 f(m+1, |P|)})$ .

Let  $q$  be the point drawn from  $Z_p$  in step 1(a) of TRIAL. We claim that  $\text{vol}(\tau_m * q) \leq \sigma_1(m+1)^{m+1} L^{m+1} / (m+1)!$ , where  $L$  is the longest edge length of  $\tau_m$ . Assume to the contrary that our claim is false. Let  $\ell$  be the longest edge length of  $\tau_m * q$ . By step 1(b) of TRIAL, we have  $\ell \leq 2r_p \leq 40L$ .

If  $\text{vol}(\tau_m * q) > \sigma_1(m+1)^{m+1} L^{m+1} / (m+1)!$ , then  $\text{vol}(\tau_m * q) > \frac{(\sigma_1(m+1)/40)^{m+1} \ell^{m+1}}{(m+1)!} \geq \frac{\sigma_2^{m+1} \ell^{m+1}}{(m+1)!}$ . For any boundary  $i$ -simplex  $\tau$  of  $\tau_m$ , the distance between  $q$  and  $\text{aff}(\tau)$  is at most  $\ell$ . So the above inequality allows us to inductively argue that  $\text{vol}(\tau * q) \geq \text{vol}(\tau_m * q) \cdot \frac{(m+1)m \dots (i+2)}{\ell^{m-i}} > \frac{(\sigma_1(m+1)/40)^{m+1} \ell^{i+1}}{(i+1)!} \geq \frac{\sigma_2^{i+1} \ell^{i+1}}{(i+1)!}$ . The last inequality holds because  $\sigma_2 = (\sigma_1(m+1)/40)^{m+1} \leq (\sigma_1(m+1)/40)^{(m+1)/(i+1)}$ . We already know that  $\tau_m$  and its boundary simplices are not  $\sigma_2$ -slivers. Then, the analysis above implies that  $\tau_m * q$  and its boundary simplices are not  $\sigma_2$ -slivers. However, this contradicts Lemma 5.11(ii). This proves our claim.

By our claim, we get  $g(|P|)$  bad simplices in step 2 of TRIAL. By our choice of  $\sigma_1(m+1)$ , one can derive that  $g(|P|) > (2e+1)e^{3/5} \sum_{p \in K} \text{Ratio}(m+1, r_p, 2\sigma_1(m+1))$  in step 2 of TRIAL. So TRIAL( $K, m+1$ ) returns FAIL.  $\square$

We are now ready to prove the main theoretical guarantees and the running time bound.

THEOREM 5.1. Let  $\mathbf{M} \subset \mathbb{R}^d$  be a manifold with dimension  $m \geq 1$  and positive reach. Let  $P$  be a set of point samples drawn from  $\mathbf{M}$  according to a Poisson process with parameter  $\lambda$ . Let  $k$  be some fixed positive integer. There is a value  $\lambda_0 = 2^{\Theta(m^6)} + 2^{\Theta(km^2)}$  such that when  $\lambda > \lambda_0$ , the dimension of  $\mathbf{M}$  can be reported correctly in  $O(kd|P|^{1+1/k})$  time with probability greater than  $1-2^{-k}$ .

*Proof.* By the Chebyshev's inequality,  $|P| \geq \lambda/2$  with probability at least  $1 - 4\lambda^{-1} > 2^{-(k+2)}$ . Thus, it holds with probability greater than  $1 - 2^{-(k+2)}$  that  $|P|$  is large enough for Lemma 5.12 and Lemma 5.13 to hold. We assume that the work zones are disjoint which happens with probability  $1 - O(k^2 m^2 2^{4m} |P|^{1/4})$  by Lemma 5.6. Since  $|P| \geq \lambda/2$ , this probability bound is much greater than  $1 - 2^{-(k+2)}$ .

The appendix shows that we can define a function  $\tilde{\sigma}_0(j, r)$  of order  $O(r^{O(j^2)})$  such that  $\sigma_0(r) = \tilde{\sigma}_0(m, r)$ . For any  $j \geq 1$ , define  $\sigma_1(j) = \frac{1}{2} \min(\tilde{\sigma}_0(j, 1/40), I(j+2)/(4e^{3/5}(2e+1)))$ ,  $f(j, |P|) = |P|^{1/(4k(j+1))}$  and  $g(|P|) = |P|^{1/4k}$ .

DIMENSION executes  $k$  instances of ESTIMATE in sequential order, each instance trying  $j = 2, 3, \dots$  until the stopping condition is met. We modify DIMENSION to run all  $k$  instances of ESTIMATE for each value of  $j$ . That is, for each  $j = 2, 3, \dots$ , the modified DIMENSION calls step 2 and step 3 of ESTIMATE  $k$  times. Each call returns a dimension estimate which is stored in an array  $A$ . If some returned estimate appears more than  $k/2$  times in  $A$ , the modified DIMENSION terminates and returns this value as the dimension of  $\mathbf{M}$ . Otherwise, the modified DIMENSION increments  $j$  and calls step 2 and

step 3 of ESTIMATE again.

A call can only underestimate when  $2 \leq j \leq m$ . We have  $\beta_0(j) = O(2^{-O(j^2)})$  and  $g(|P|) = |P|^{1/(4k)}$ . Because  $|P| = \Omega(2^{\Theta(km^2)})$ , we get  $e^{-\beta_0(j)g(|P|)} = e^{-\Omega(2^{\Theta(m^2)})}$  which is much less than  $1/(16m^2)$ .

Before  $j$  exceeds  $m$ , we cannot overestimate the dimension of  $M$ . Consider  $j = m + 1$ . Given the definitions of  $f(j, |P|)$  and  $g(|P|)$ , the probability bound in Lemma 5.13 becomes  $O(80^m |P|^{-1+1/(4k)} + 80^m |P|^{1/(4k)} e^{-\beta_1 |P|^{1/(4k(m+2))}})$ . For  $k \geq 1$ , the first term is  $O(80^m |P|^{-3/4})$  which is much less than  $1/32$  because  $|P| = \Omega(2^{\Theta(m^6)})$ . Consider the second term. Since  $\beta_1 = 80^{-m} e^{-1/2}/32$  and  $|P|^{1/(4k(m+2))} = \Omega(2^{\Theta(m)})$ , we have  $\frac{1}{2}\beta_1 |P|^{1/(4k(m+2))} > \frac{1}{2}|P|^{1/(8k(m+2))} > \ln(80^m |P|^{1/(4k)})$ . It follows that  $e^{-\frac{1}{2}\beta_1 |P|^{1/(4k(m+2))}}$  cancels  $80^m |P|^{1/(4k)}$ . This leaves  $O(e^{-\frac{1}{2}\beta_1 |P|^{1/(4k(m+2))}}) = O(e^{-2^{\Omega(m)}})$  which is much less than  $1/32$ . Thus, a call overestimates with probability less than  $1/16$ .

The modified DIMENSION may give a wrong answer if one or more of the following happen: (i)  $|P| < \lambda/2$  which happens with probability less than  $2^{-(k+2)}$ ; (ii) overlapping work zones which happens with probability less than  $2^{-(k+2)}$ ; (iii)  $k/2$  or more calls underestimate for some  $2 \leq j \leq m$ , which happens with probability less than  $(m-1) \cdot (16^{-k/2} m^{-k}) < 2^{-2k}$ ; (iv)  $k/2$  or more calls overestimate when  $j = m + 1$ , which happens with probability less than  $16^{-k/2} \leq 2^{-2k}$ . In all, the failure probability is less than  $2^{-(k+2)} + 2^{-(k+2)} + 2^{-2k} + 2^{-2k} < 2^{-k}$  as desired.

Since we stop at  $j = m + 1$  with probability greater than  $1 - 2^{-k}$ , by Lemma 4.1, the running time in this case is  $O(kdm |P|^{1+1/(4k)} + kdm^3 |P|^{1/(2k)} + m^3 |P|^{1/(2k)} \log |P|)$ . Because  $|P| = \Omega(2^{\Theta(km^2)})$ , we have  $m < m^3 < m^4 = O(|P|^{1/(2k)})$ . Clearly,  $\log |P| = O(|P|^{1/(2k)})$ . Thus, the running time can be simplified to  $O(kd |P|^{1+1/k})$ .  $\square$

*Remark 1.* In the worst case, the modified DIMENSION only stops at  $j = d$ . By Lemma 4.1, the worst-case running time is  $O(kd^2 |P|^{1+1/k} + kd^4 |P|^{1/k})$ .

*Remark 2.* The dimension of  $M$  can also be estimated with high probability by setting  $f(j, n) = g(n) = \Theta(\ln n)$ . This is a simple way to get a reasonable number of neighbors to work with in the experiments. However, the theoretical analysis goes through only if  $\lambda$  is doubly exponential in some polynomial in  $m$  then.

## 6 Experiments.

We implemented a practical variant of our algorithm. that incorporates some heuristical improvements. Although we sacrifice the theoretical guarantees by doing

so, the practical variant performs well in the experiments. It does not require a high sampling density and it is competitive with several previous methods.

**6.1 Practical variant.** We set  $f(j, |P|) = 2.5 \ln |P|$  for all  $j$ . We have between 50 and 7000 points in our experiments and this gives between 11 to 24 points in a work zone. We also fix the number  $k$  of for-loops in DIMENSION to be 25.

A lot of time is spent in steps 1(b) and 1(c) of TRIAL in finding a  $(j-1)$ -simplex  $\tau_{j-1}$  just to check if we can obtain one single bad simplex  $\tau_{j-1} * q$ . This is rather wasteful. Thus, we modify ESTIMATE such that  $K$  contains only one random trial point  $p$ . So ESTIMATE calls TRIAL( $\{p\}, j$ ) for  $j = 2, 3, \dots$ . We also modify TRIAL. We do not draw a random point  $q$  from  $Z_p$  in step 1(a). If a  $(j-1)$ -simplex has volume very close to  $\sigma_1^{j-1} L^{j-1} / (j-1)!$ , it is somewhat arbitrary to call it a sliver or non-sliver. Thus, we use two values  $(\sigma_1, \sigma_2)$ , where  $\sigma_1 < \sigma_2$ . In step 1(b), we construct a list  $L_{p,j}$  of all  $(j-1)$ -simplices  $\tau_{j-1}$  such that: (i)  $\tau_{j-1}$  has  $p$  and other points in  $Z_p$  as vertices; (ii) the shortest edge length of  $\tau_{j-1}$  is at least  $r_p/20$ ; (iii)  $\tau_{j-1}$  and its boundary simplices are not  $\sigma_2$ -slivers. In step 1(c), for each simplex  $\tau_{j-1}$  in  $L_{p,j}$  and for each point  $q \in Z_p$  that is not a vertex of  $\tau_{j-1}$ , we call  $\tau_{j-1} * q$  bad if it is a  $\sigma_1$ -sliver. We accumulate the number of bad simplices in the variable  $N_{p,j}$ . Motivated by Lemma 5.8, we expect to encounter roughly  $\Delta_{p,j} = |L_{p,j}| \cdot (2.5 \ln |P| + 1 - j) \cdot \text{Ratio}(j, r_p, 2\sigma_1)$  bad simplices. The advantage of using a single work zone  $Z_p$  is that  $L_{p,j}$  computed in the call TRIAL( $\{p\}, j$ ) can be used to speed up the construction of  $L_{p,j+1}$  in the call TRIAL( $\{p\}, j+1$ ).

The remaining issue is how to pick  $(\sigma_1, \sigma_2)$  and how to terminate ESTIMATE. Suppose that  $(\sigma_1, \sigma_2)$  is the best choice for a dataset. If we increase  $\sigma_1$  and  $\sigma_2$ , it becomes hard for  $L_{p,j}$  to be non-empty in the call TRIAL( $\{p\}, j$ ). Conversely, if we decrease  $\sigma_1$  and  $\sigma_2$ , it becomes hard to classify  $\tau_{j-1} * q$  as a bad simplex. It becomes harder as  $j$  increases because  $\sigma_1^j$  drops and  $\text{vol}(\tau_{j-1} * q)$  needs to be at most  $\sigma_1^j L^j / j!$  for  $\tau_{j-1} * q$  to be bad. This tends to cause an underestimation. Hence, we run DIMENSION (after modifying ESTIMATE and TRIAL as described previously) three times with  $(\sigma_1, \sigma_2) = (0.6, 0.7), (0.5, 0.6)$  and  $(0.4, 0.5)$  in this order. ESTIMATE terminates when the calling of TRIAL( $\{p\}, j$ ) meets one of the following conditions: (i)  $N_{p,j} > \Delta_{p,j}$ ; (ii)  $L_{p,j}$  is empty; (iii) no bad simplex is identified and  $j$  exceeds the highest output of the previous runs of DIMENSION using some other values of  $(\sigma_1, \sigma_2)$ . Condition (i) is the analog of step 2 in the original TRIAL, but we need the other conditions because  $N_{p,j}$  may not exceed  $\Delta_{p,j}$  in practice. If condition (ii) holds,  $\sigma_2$  is probably too large



Figure 1: Head.



Figure 2: S1 and S0.

or  $j$  has probably exceeded the manifold dimension. We stop on condition (iii) because  $\sigma_1$  is probably too small and  $j$  has become so large that we cannot find a  $j$ -simplex that is a  $\sigma_1$ -sliver.

When ESTIMATE stops at  $j$ , we find  $i \in [2, j]$  that maximizes  $N_{p,i}/\Delta_{p,i}$  and return  $i - 1$ . In the experiments, each run of DIMENSION makes 25 calls to ESTIMATES and output the value of the highest frequency among the returned ones. Finally, we pick the highest output among the three runs of DIMENSION as the final answer. We observe in the experiments that, as  $i$  increases, the ratio  $N_{p,i}/\Delta_{p,i}$  increases to a maximum and then drops monotonically.

**6.2 Results.** We did the experiments on a PC with an Intel Core 2 CPU 6400 2.13GHz and 0.99GB RAM. We compared our program with the maximum likelihood estimation (MLE) [16], the manifold adaptive method (MA) [10], the packing number method (PN) [15], and the local PCA (LPCA) [5]. The LPCA was invoked on the neighborhoods determined by our practical variant. We also ran ISOMAP [20], which is a manifold embedding method, on some datasets. It returns a plot of the residual variance as the target dimension. Abrupt flattening in the plot indicates the manifold dimension. ISOMAP is often used in the literature for comparison. We experimented with seven datasets. They include points uniformly distributed on  $n$ -dimensional unit spheres for  $4 \leq n \leq 9$ , 698 images of a rotating head (Head, Fig. 1), 2240 images of 2D translations of a smaller image (Shift), 5923 images of synthetic zeros (S0), 6742 images of synthetic ones (S1),



Figure 3: H1 and H0.

images of handwritten ones (H1, Fig. 3) and zeros (H0, Fig. 3) from the MNIST database [21].

Table 1 shows the results for the sphere datasets. For each  $n$  and for each dataset size, we tried each method on 30 sets of the same size and recorded the number of successes. Our program performs the best. Table 2 shows the results on the other datasets. All methods give an estimate close to 3 or 4 for Head, which is consistent with the dimension reported in the literature. Shift has dimension 2 and ISOMAP outputs 2. Our programs and LPCA output 3. The others overestimated by more. No ground truth is known for H1 and H0, so we first try some synthetic ones (S1) and zeros (S0). S1 consists of segments with different length, width and rotation. So the dimension is 3. S0 consists of ellipses with different standard ellipse parameters, width, and rotation. So the dimension is 4. The best answers for S1 are given by our program and PN. PN and LPCA perform the best for S0 followed by our program. ISOMAP gives the ranges [2,3] and [2,4] for S1 and S0, respectively. We believe that the dimensions of H1 and H0 lie in the ranges [3,7] and [2,9], respectively, which are the ranges of dimensions output by our program, PN, LPCA, and ISOMAP.

Table 1: The numbers of successes in each entry are in this order Ours,MLE,MA,PN,LPCA.

	50 pts	100 pts	500 pts
$S^4$	28,30,21,6,1	30,30,29,6,5	30,30,30,9,23
$S^5$	29,23,12,0,0	27,30,21,0,0	30,30,30,0,0
$S^6$	30,7,5,0,0	29,23,1,0,0	30,30,30,0,0
$S^7$	22,0,1,0,0	30,8,1,0,0	30,30,29,0,0
$S^8$	2,0,0,0,0	27,2,0,0,0	30,30,9,0,0
$S^9$	0,0,0,0,0	9,0,0,0,0	30,18,2,0,0
	1000 pts	3000 pts	5000 pts
$S^4$	30,30,30,13,30	30,30,30,25,30	30,30,30,27,30
$S^5$	30,30,30,0,6	30,30,30,0,9	30,30,30,2,20
$S^6$	30,30,20,0,0	30,30,30,0,0	30,30,30,0,0
$S^7$	29,30,0,0,0	29,30,30,0,0	30,30,7,0,0
$S^8$	30,30,0,0,0	30,30,30,0,0	30,30,0,0,0
$S^9$	30,30,0,0,0	30,30,2,0,0	30,30,0,0,0

Table 2: Six other datasets.

	Head	Shift	S1	S0	H1	H0
Ours	4	3	3	2	3	2
MLE	4.31	4.27	5.11	6.88	11.47	14.86
MA	4.47	3.35	6.38	7.26	10.77	13.93
PN	3.98	3.62	2.53	5.14	6.22	8.86
LPCA	3	3	5	5	5	7
ISOMAP	3	2	[2,3]	[2,4]	5	[3,6]

## 7 Conclusion

We presented an algorithm and a practical variant to estimate the dimension of a manifold. The experimental results suggest that the lower bound of  $\Omega(2^{\Theta(m^6)} + 2^{\Theta(km^2)})$  on  $\lambda$  may not be tight.

## 8 Acknowledgment

We thank the anonymous referees for helpful comments.

## References

- [1] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems*, 14 (2002). MIT Press.
- [2] J.-D. Boissonnat, L.J. Guibas and S.Y. Oudot. Manifold reconstruction in arbitrary dimensions using witness complexes. *Proc. 23rd ACM Sympos. Comput. Geom.*, 2007, 194–203.
- [3] S.-W. Cheng, T.K. Dey, H. Edelsbrunner, M.A. Facello, and S.-H. Teng. Sliver Exudation. *J. ACM*, 47 (2000), 883–904.
- [4] S.-W. Cheng, T.K. Dey and E.A. Ramos. Manifold reconstruction from point samples. *Proc. 16th Annu. ACM-SIAM Sympos. Discrete Alg.*, 2005, 1018–1027. Journal version in preparation, 2008.
- [5] S.-W. Cheng, Y. Wang and Z. Wu. Provable dimension detection using principal component analysis. To appear in *Int'l. J. Comput. Geom. Appl.* Preliminary version appears in *Proc. 21st Annu. Sympos. Comput. Geom.*, 2005, 208–217.
- [6] J. A. Costa and A. O. Hero. Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *IEEE Transactions on Signal Processing*, 52(8):2210–2221, 2004.
- [7] C.D. Cutler. A review of the theory and estimation of fractal dimension. In *Dimension Estimation and Models*, (edited by H. Tong), World Scientific, 1993.
- [8] T. K. Dey, J. Giesen, S. Goswami and W. Zhao. Shape dimension and approximation from samples. *Discr. Comput. Geom.*, 29 (2003), 419–434.
- [9] D. Donoho and C. Grimes. Hessian eigenmaps: locally linear embedding techniques for high dimensional data. *Proc. National Academy of Sciences*, 100(10):5591–5596, 2003.
- [10] A. massoud Farahmand, C. Szepesvári and J.-Y. Audibert. Manifold-adaptive dimension estimation. *Proc. 24th Int'l. Conf. Machine Learning*, 2007, 265–272.
- [11] J. Giesen and U. Wagner. Shape dimension and intrinsic metric from samples of manifolds with high codimension. *Discr. Comput. Geom.*, 32 (2004), 245–267.
- [12] G.H. Golub and C.F. van Loan. *Matrix Computations*, 3rd edition, Johns Hopkins, 1996.
- [13] P. Grassberger and I. Procaccia. Measuring the strangeness of strange attractors. *Physica D*, 9 (1983), 189–208.

- [14] M. Hein and J.-Y. Audibert. Intrinsic dimensionality estimation of submanifolds in Euclidean space. *Proc. 22nd Int'l. Conf. Machine Learning*, 2005, 289–296.
- [15] Balázs Kégl. Intrinsic dimension estimation using packing numbers. In *Advances in Neural Information Processing Systems 15*, 14 (2003), 681–688. MIT Press.
- [16] E. Levina and P.J. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems 17*, 777–784, 2005. MIT Press.
- [17] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.
- [18] K.W. Pettis, T.A. Bailey, A.K. Jain, and R.C. Dubes. An intrinsic dimensionality estimator from near neighbor information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1 (1979), 25–37.
- [19] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. In *Science*, volume 290, pages 2323–2326, 2000.
- [20] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.
- [21] <http://www.cs.toronto.edu/~roweis/data.html>

## A Appendix

We show that  $\sigma_0(\delta/\varepsilon) = O((\delta/\varepsilon)^{O(m^2)})$ . It is known from the analysis in [4] that  $\sigma_0(\delta/\varepsilon) = O((\frac{\delta}{\varepsilon})^2 \cdot (\sum_{i=2}^{m+1} \binom{N}{i+1})^{-1})$ , where  $N$  is the total number of restricted weighted Delaunay edges incident a point  $p$  over all possible weighted Delaunay triangulation with weight property  $[\omega]$  for some  $\omega < 1/2$ . So it suffices to prove that  $N = O((\varepsilon/\delta)^m)$ .

Assume that  $p$  has unit nearest neighbor distance. It has been shown in [4] that: (i) the longest weighted Delaunay edge incident to  $p$  has length at most  $\nu = O(\varepsilon/\delta)$ ; (ii) any restricted Delaunay edge incident to  $p$  makes an angle  $\Theta(\varepsilon)$  with  $T_p(\mathbb{M})$ ; (iii) for any restricted weighted Delaunay edge  $pq$ , the nearest neighbor distance of  $q$  is at least  $1/\nu$ .

By (iii), for each restricted weighted Delaunay edge  $pq$ , we can assign a ball  $B(q, 1/(2\nu))$  and these balls are disjoint. By (ii), each such  $q$  is at distance  $O(\varepsilon\nu)$  from  $T_p(\mathbb{M})$ . This implies that the cross-section  $B(q, 1/(2\nu)) \cap T_p(\mathbb{M})$  is a  $m$ -dimensional ball with volume  $\Theta((2\nu)^{-m})$ . These  $m$ -dimensional cross-sections are disjoint and each lies inside the  $m$ -dimensional ball in  $T_p(\mathbb{M})$  centered at  $p$  with radius  $\nu + 1/(2\nu)$ . Thus, a packing argument shows that there are  $O((2\nu + 1)^m) = O((\varepsilon/\delta)^m)$  cross-sections. Hence,  $N = O((\varepsilon/\delta)^m)$ .