Project number IST-25582

CGL

Computational Geometric Learning

"Sketching the Support of a Probability Measure"

STREP

Information Society Technologies

Period covered:	November 1, 2012–October 31, 2013
Date of preparation:	October 31, 2013
Date of revision:	October 31, 2013
Start date of project:	November 1, 2010
Duration:	3 years
Project coordinator name:	Joachim Giesen (FSU)
Project coordinator organisation:	Friedrich-Schiller-Universität Jena Jena, Germany

Sketching the Support of a Probability Measure

Joachim Giesen Lars Kühne Sören Laue

November 4, 2013

Abstract

We want to sketch the support of a probability measure on Euclidean space from samples that have been drawn from the measure. This problem is closely related to certain manifold learning problems, where one assumes that the sample points are drawn from a manifold that is embedded in Euclidean space. Here we propose to sketch the support of the probability measure (that does not need to be a manifold) by some gradient flow complex, or more precisely by its Hasse diagram. The gradient flow is defined with respect to the distance function to the sample points. We prove that a gradient flow complex (that can be computed) is homotopy equivalent to the support of the measure for sufficiently dense samplings, and demonstrate the feasibility of our approach on real world data sets.

1 Introduction

Our goal is to compute a sketch, i.e., an approximation, of the support $\operatorname{supp}(\mu)$ of a probability measure μ on $\mathbb{R}^d$ from a finite set of sample points that are drawn from μ. More specifically, we want to compute a complex (not necessarily geometrically realized) that has the same homotopy type as $\operatorname{supp}(\mu)$.

Formally, a probability measure and its support are defined as follows.

Probability measure. A non-negative measure μ on $\mathbb{R}^d$ is an additive function that maps every Borel subset $B \subseteq \mathbb{R}^d$ to $\mathbb{R}_{\geq 0}$. Additivity means that

$$\mu\left(\bigcup_{i\in\mathbb{N}} B_i\right) = \sum_{i\in\mathbb{N}} \mu(B_i),$$

where (B_i) is a countable family of disjoint Borel subsets. The measure μ is finite if $\mu(\mathbb{R}^d) < \infty$ and it is a probability measure if $\mu(\mathbb{R}^d) = 1$.

The support of a probability measure μ is the set

$$\operatorname{supp}(\mu) = \{x \in \mathbb{R}^d \mid \mu(B(x,r)) > 0 \text{ for all } r > 0\},$$

where $B(x,r)$ is the closed ball with radius r that is centered at x. Note that $\operatorname{supp}(\mu)$ is always closed.

Given samples $x_1, \ldots, x_n$ drawn from μ, a natural approach to sketch $\operatorname{supp}(\mu)$ is to approximate it with a union of balls centered at the sample points, i.e., by

$$X_\alpha = \bigcup_{i=1}^{n} B(x_i, \alpha),$$

where $B(x_i, \alpha)$ is the ball of radius $\alpha > 0$ centered at x_i. The obvious problem with this approach is to determine a good value for α. One contribution of our paper is a simple method for choosing such a good value for α.

The homotopy type of the union of balls X_α can be computed from the nerve of the ball covering, i.e., the simplicial complex that is determined by the intersection pattern of the balls, see for example [11] for an introduction to computational topology. The latter complex is known as the Cech complex of the union. A smaller simplicial complex from which the homotopy type of the union of balls can also be computed is the α-complex which is the nerve of the ball covering where the balls are restricted to the Voronoi cells of their centers. Yet another complex from which the homotopy type can be computed is the α-flow complex. This complex is even smaller than the α-complex, i.e., does contain a smaller number of cells, but it is no longer simplicial. The flow complex can be derived from the distance function to the sample points. It contains a cell for every critical point of the distance function. We show that the critical points of the distance function are either close to $\mathrm{supp}(\mu)$ or close to a dual structure of $\mathrm{supp}(\mu)$ that is called the medial axis of $\mathrm{supp}(\mu)$. If the support of μ does not exhibit several geometric scales, then the critical points that belong to $\mathrm{supp}(\mu)$ and the critical points that belong to the medial axis can be separated by simple thresholding, i.e., all critical points at which the distance function takes values less than the threshold value belong to $\mathrm{supp}(\mu)$ and the remaining critical points belong to the medial axis. Restricting the flow complex to the critical points with distance function values less than α constitutes the α-flow complex. If α is a threshold value at which the two types of critical points can be separated, then the α-flow complex is homotopy equivalent to $\mathrm{supp}(\mu)$ for sufficiently dense samplings.

We have computed α-flow complexes for real data sets and all values for α in the interval $[0, \infty)$. On these data sets we have not observed geometric multi-scale behavior. Hence, in these situations the simple thresholding was enough to compute a complex that is homotopy equivalent to $\mathrm{supp}(\mu)$ for sufficiently dense samplings. We report details on the data sets and our findings at the end of this paper.

Related work. Our work is related to certain manifold learning problems that we briefly summarize here. In machine learning manifold learning is often used synonymously with non-linear dimensionality reduction, but there is also quite some work (mostly in computational geometry) that aims at learning a manifold from samples (that need to satisfy certain conditions), where learning a manifold refers to computing an approximation from a finite sampling that is guaranteed to be topologically equivalent and geometrically close to the manifold. Exemplary for this line of work is the technique by Boissonnat and Ghosh [3]. The body of work in computational geometry does not consider the probabilistic setting where the sample points are drawn at random from the manifold. The probabilistic setting was first considered by Niyogi et al. [19] who show how to compute the homology of a randomly sampled manifold with high confidence. Later Niyogi et al. [20] have extended this approach for recovering the geometric core of Gaussian noise concentrated around a low dimensional manifold, i.e., to the case where the samples are not necessarily drawn from the manifold itself. This can be seen as a topological approach to unsupervised learning.

Manifold learning plays an important role in semi-supervised classification, where a *manifold assumption*, see [1], can be used to exploit the availability of unlabeled data in classification tasks. The assumption requires that the support of the marginal probability distribution underlying the data is a manifold (or close to a manifold). The manifold assumption for semi-supervised learning has been exploited by Belkin et al. [1] in form of a support vector machine with an additional Laplacian regularization term (Laplacian SVM), see also [17]. For Laplacian SVMs the manifold is approximated just by some neighborhood graph on the data points that can be computed efficiently also in high dimensions, but does not come with

approximation guarantees. Laplacian SVMs have been shown to achieve state of the art performance in semi-supervised classification.

Our approach here is more abstract. We are also in the probabilistic setting, but do not assume that the support of the probability measure from which the samples are drawn is a manifold. Still, also in this setting we can provide topological reconstruction guarantees.

2 Distance Function

Here we briefly review the theory of distance functions to a compact set that has been developed within the fields of differential and computational geometry over the last years [14, 10, 22, 12, 9, 16, 8, 13, 4]. In the following let K always denote a compact subset of $\mathbb{R}^d$.

Distance function. The distance function d_K to the compact set K assigns to any point $x \in \mathbb{R}^d$ its distance to K, i.e.,

$$d_K : \mathbb{R}^d \to [0, \infty), x \mapsto \min_{y \in K} \|x - y\|.$$

The function d_k characterizes K completely since $K = d_K^{-1}(0)$.

Gradient. For any point $x \in \mathbb{R}^d$ let

$$N_K(x) = \{p \in K \ : \ \|x - p\| = d_K(x)\}$$

be the set of nearest neighbors in K and let $c(x)$ be the center of the smallest enclosing ball of $N_K(x)$. The gradient of the distance function at x is given as

$$\partial_K(x) = \frac{x - c(x)}{d_K(x)}, \quad \text{if} \quad x \neq c(x),$$

and 0 otherwise. The norm of the gradient is always upper bounded by 1, i.e., $\|\partial_K(x)\| \leq 1$.

Medial axis. The medial axis of K is the following set

$$\mathrm{ma}(K) \ = \ \big\{x \in \mathbb{R}^d \setminus K \ \big| \ \|\partial_K(x)\| < 1\big\},$$

i.e., the set of all center points of maximal empty open balls in the complement $\mathbb{R}^d \setminus K$ of K.

Reach. The reach of K is defined as

$$\inf_{x \in K, \, y \in \mathrm{ma}(K)} \|x - y\|.$$

If the reach of K is positive, then we say that K has finite reach.

The approximation guarantees for our sketch, i.e., preserving the homotopy type, are based on a reconstruction theorem for compact subsets of Euclidean space that has been proved in [5]. This theorem topologically relates the off-sets of two compact sets that need to satisfy certain conditions.

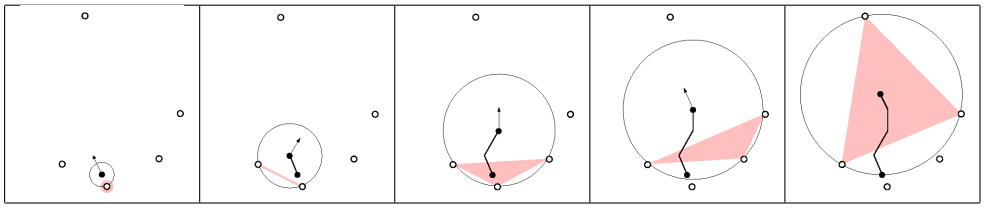

Figure 1: Example flow line in two dimensions for a set X of five sample points. Shown on the left is the starting point x of the flow line together with the starting flow direction that is determined by the smallest enclosing ball of $N_X(x)$, which here is just a sample point. Then, shown from left to right are the points where the set $N_X(x)$, and thus the direction of flow, changes. The flow line ends (shown on the right) in a local maximum of the distance function, i.e., x is contained in the convex hull of $N_X(x)$.

α-offset. For any set $K \subset \mathbb{R}^d$ and $\alpha > 0$ let K_α be the Minkowski sum of K and $B(0, \alpha)$, i.e.,

$$K_\alpha = \left\{ x \in \mathbb{R}^d \,|\, x \in B(x', \alpha),\, x' \in K \right\}.$$

Now we are equipped with the necessary definitions to state the reconstruction theorem.

Theorem 1. [Chazal et al.] *Let $\rho > 0$ be the reach of K and let $K' \subset \mathbb{R}^d$ be a compact set such that the Hausdorff distance between K and K' is less than $\frac{\rho}{17}$, i.e., $d_H(K, K') < \frac{\rho}{17}$, then the complement $\mathbb{R}^d \setminus K'_\alpha$ of K'_α is homotopy equivalent to the complement $\mathbb{R}^d \setminus K$ of K, and K'_α is homotopy equivalent to K_η for all sufficiently small $\eta > 0$, provided that*

$$4 \cdot d_H(K, K') \leq \alpha \leq \rho - 3 \cdot d_H(K, K').$$

□

As we have already indicated in the introduction, often K' is a finite sampling of K. Let X be such a finite sampling, then $X_\alpha = K'_\alpha$ is a union of balls with radius α that is homotopy equivalent to K_η for all sufficiently small $\eta > 0$, if the Hausdorff distance between X_α and K is small, and α is in the range given by Theorem 1. Note that the practical problem of choosing a good value for α, i.e., a value that falls into the range specified by Theorem 1 still remains open. In the following we address this problem by considering the critical points of the distance function to the set $X \subset K$.

3 Flow Complex

The flow complex of a finite point set $X \subset \mathbb{R}^d$ is a cell complex that contains a cell for every critical point of the distance function to X. We show that a properly chosen sub-complex of the flow complex of X, namely some α-flow complex, provides a homotopy equivalent reconstruction of the support of a probability measure μ given that X is a sufficiently dense sampling drawn from μ.

Critical points. Let K be a compact subset of $\mathbb{R}^d$. The points $x \in \mathbb{R}^d$ with $\partial_K(x) = 0$, i.e., the points for which $x = c(x)$, are called the critical points of the distance function d_K (cf. the definition of the gradient of d_K in the previous

section). Critical points x of d_K are always contained in the convex hull of their neighbors in K, i.e., in $\mathrm{conv}(N_K(x))$. Note that all the critical points of the distance function d_K are either points of K, or are contained in the medial axis $\mathrm{ma}(K)$.

If K is a finite point set X, then a meaningful index $i(x)$ can be assigned to any critical point x of d_X, namely the dimension of the affine hull of $N_X(x)$. Critical points of index 0 are the points in X, i.e., the minima of the distance function. Critical points of index d are maxima of d_X, and all other critical points are saddle points of d_X. All critical points with positive index are contained in $\mathrm{ma}(X)$.

Flow complex. Let $X \subset \mathbb{R}^d$ be a finite point set. The flow induced by the gradient vector field ∂_X is a mapping

$$\phi_X : [0, \infty) \times \mathbb{R}^d \to \mathbb{R}^d$$

defined by the equations $\phi_X(0, x) = x$ and

$$\lim_{t \downarrow t_0} \frac{\phi_X(t, x) - \phi_X(t_0, x)}{t - t_0} \;=\; \partial_X(\phi_X(t_0, x)).$$

The set $\phi_X(x) = \{\phi_X(t, x) \,|\, t \geq 0\}$ is called the flow line of the point x, see Figure 1. The stable manifold $S(x)$ of a critical point x is the set of all points in $\mathbb{R}^d$ that flow into x, i.e.,

$$S(x) = \{y \in \mathbb{R}^d \;:\; \lim_{t \to \infty} \phi_X(t, y) = x\}.$$

The dimension of $S(x)$ is given by the index $i(x)$. The flow complex is given by the stable manifolds of all critical points together with the following incidence information that is defined using the unstable manifolds of critical points. Given a neighborhood U of a critical point x and setting

$$V(U) \;= \\ \{y \in \mathbb{R}^d \,|\, \exists z \in U, t \geq 0 \text{ such that } \phi_X(t, z) = y\},$$

the unstable manifold of x is the set

$$U(x) \;= \bigcap_{\text{neighborhood } U \text{ of } x} V(U).$$

The stable manifold of a critical point y is incident to the stable manifold of a critical point x if

$$S(x) \cap U(y) \neq \emptyset,$$

i.e., if there is a point in the unstable manifold of y that flows into x. The incidence structure on the stable manifolds of the critical points is a binary relation that is

1. *reflexive*, because $S(x) \cap U(x) = \{x\}$ for any critical point x.
2. *antisymmetric*, because $S(x) \cap U(y) \neq \emptyset$ and $S(y) \cap U(x) \neq \emptyset$ implies $x = y$.
3. *transitive*, because $S(x) \cap U(y) \neq \emptyset$ implies $U(x) \subseteq U(y)$, and hence if x is incident to z, i.e., $S(z) \cap U(x) \neq \emptyset$, then also y is incident to z.

Hence, the combinatorial structure of the flow complex induces a partial order on the set of stable manifolds, or the critical points of d_X, respectively, that can be encoded in a Hasse diagram, see Figure 2.

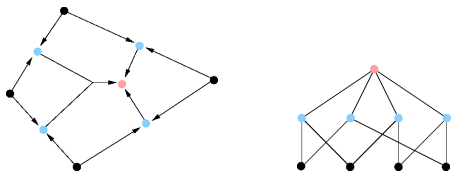

Figure 2: On the left: A finite point set of four points (in black) in two dimensions together with the critical points of its distance function (index-0 (black), index-2 (light blue), and index-2 (pink)). The arrows represent flow lines that witness the incidence relationships between the critical points. On the right: The Hasse diagram of the flow complex.

α-flow complex. For $\alpha \geq 0$, the α-flow complex of of a finite point set $X \subset \mathbb{R}^d$ is the Hasse diagram of the flow complex restricted to the critical points x of d_X for which $d_X(x) \leq \alpha$.

In [7] it has been shown that the union of balls X_α and the α-flow complex of the finite point set X are homotopy equivalent. Hence, also the α-flow complex of X is homotopy equivalent to K_η for small $\eta > 0$, if the Hausdorff distance between X_α and K is small, and α is in the range given by Theorem 1.

4 Topological Guarantees

In this section we specify conditions under which an α-flow complex of the sample points in $X = \{x_1, \ldots, x_n\}$ drawn from μ is homotopy equivalent to $\text{supp}(\mu)$. We do so by using the following lemma, see [6](Lemma 5.1).

Lemma 2. *Given a sequence of sample points $x_1, \ldots, x_n$ drawn independently from a probability measure μ on $\mathbb{R}^d$. Then, for every $\varepsilon > 0$ and any $x \in \text{supp}(\mu)$,*

$$\lim_{n\to\infty} P\big[\|x_1(x) - x\| > \varepsilon\big] = 0,$$

where $x_1(x)$ is the nearest neighbor of x in $\{x_1, \ldots, x_n\}$. □

An immediate consequence of this lemma is the following corollary.

Corollary 3. *Given a sequence of sample points $x_1, \ldots, x_n$ drawn independently from a probability measure μ with compact support on $\mathbb{R}^d$. Then, for every $\varepsilon > 0$,*

$$\lim_{n\to\infty} P\big[d_H\big(\text{supp}(\mu),\, X\big) > \varepsilon\big] = 0,$$

where $d_H\big(\text{supp}(\mu),\, X\big)$ is the Hausdorff distance between $\text{supp}(\mu)$ and $X = \{x_1, \ldots, x_n\}$.

Proof. Since $\text{supp}(\mu)$ is compact there is a finite set of points $y_1, \ldots, y_m \in \text{supp}(\mu)$ such that the union of balls $\bigcup_{i=1}^m B(y_i, \varepsilon/2)$ covers $\text{supp}(\mu)$. Assume there exists $y \in \text{supp}(\mu)$ such that $\min_{x\in X} \|x - y\| > \varepsilon$. By construction there exists y_i such

that $\|y - y_i\| \leq \varepsilon/2$, and thus $\min_{x \in X} \|x - y_i\| > \varepsilon/2$. It follows that

$$\begin{aligned} & P\left[\sup_{y \in \operatorname{supp}(\mu)} \min_{x \in X} \|x - y\| > \varepsilon\right] \\ & \leq P\left[\max_{y \in \{y_1, \ldots, y_m\}} \min_{x \in X} \|x - y\| > \varepsilon/2\right] \\ & \leq \sum_{i=1}^{m} P\left[\min_{x \in X} \|x - y_i\| > \varepsilon/2\right] \\ & = \sum_{i=1}^{m} P\left[\min_{x \in X} \|x_1(y_i) - y_i\| > \varepsilon/2\right], \end{aligned}$$

where the second inequality follows from a simple union bound. From Lemma 2 we have

$$\lim_{n \to \infty} P[\|x_1(y_i) - y_i\| > \varepsilon/2] = 0,$$

for all y_i, and thus

$$\lim_{n \to \infty} P\left[\sup_{y \in \operatorname{supp}(\mu)} \min_{x \in X} \|x - y\| > \varepsilon\right] = 0,$$

which implies the claim on the Hausdorff distance since we also have $x_i \in \operatorname{supp}(\mu)$ for all sample points and thus

$$\max_{x \in \{x_1, \ldots, x_n\}} \inf_{y \in \operatorname{supp}(\mu)} \|x - y\| = 0.$$

□

Now we are prepared to state and prove our topological approximation guarantees.

Theorem 4. *Given a sequence of sample points $x_1, \ldots, x_n$ drawn independently from a probability measure μ with compact support on $\mathbb{R}^d$ whose reach ρ is positive. Then, for every $0 < \alpha < \rho$ and sufficiently small $\eta > 0$,*

$$\lim_{n \to \infty} P\big[\ \textit{the } \alpha\textit{-flow complex of } \{x_1, \ldots, x_n\} \textit{ is not homotopy equivalent to } \textit{supp}_\eta(\mu)\ \big] = 0.$$

Proof. Since the α-flow complex of $X = \{x_1, \ldots, x_n\}$ is homotopy equivalent to the union of balls $B(x_i, \alpha), i = 1, \ldots, n$ it suffices to show that

$$\lim_{n \to \infty} P\left[\bigcup_{i=1}^{n} B(x_i, \alpha) \text{ is not homotopy equivalent to } \operatorname{supp}_\eta(\mu)\right] = 0.$$

For that we check that α satisfies the conditions of the reconstruction theorem (Theorem 1). By Corollary 3,

$$\lim_{n \to \infty} P\big[d_H\big(\operatorname{supp}(\mu),\, X\big) > \varepsilon\big] = 0$$

for every $\varepsilon > 0$. Hence,

$$\lim_{n \to \infty} P\big[4 \cdot d_H\big(\operatorname{supp}(\mu), X\big) > \alpha\big] = 0,$$

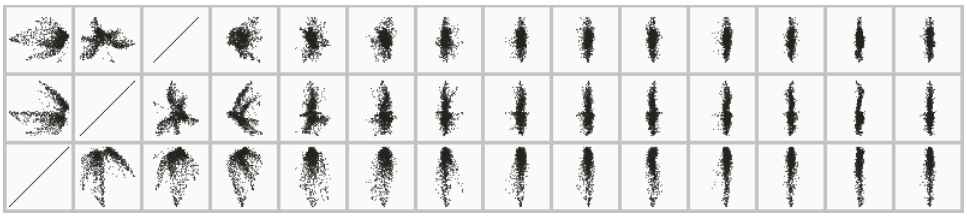

Figure 3: First three rows of the scatter plot matrix of the 14-dimensional embedding of the MovieLens data set.

and

$$\lim_{n\to\infty} P\big[\rho - 3\cdot d_H\big(\mathrm{supp}(\mu), X\big) < \alpha\big] = 0,$$

which implies the claim about the homotopy equivalence of $\bigcup_{i=1}^n B(x_i,\alpha)$ and $\mathrm{supp}_\eta(\mu)$ and hence the claim of the theorem. □

5 Choosing a good value for α

In this section we prove a theorem that allows to chose a good value for α in practice. The theorem states that the critical points of the distance function to the set $X = \{x_1,\ldots,x_n\}$ of sample points can be partitioned into two subsets. The first set contains the critical points that are close to $\mathrm{supp}(\mu)$, and the second set contains the critical points that are close to the medial axis $\mathrm{ma}(\mu)$ of $\mathrm{supp}(\mu)$, i.e., there are no critical points in the complement of $\mathrm{supp}(\mu)\cup\mathrm{ma}(\mu)$ or more precisely in

$$\begin{aligned}&\mathrm{compl}_\varepsilon(\mu)\\ &\quad= \mathrm{closure}\big(\mathrm{conv}(\mathrm{supp}(\mu)) \setminus \big(\mathrm{supp}_\varepsilon(\mu)\cup\mathrm{ma}_\varepsilon(\mu)\big)\big)\end{aligned}$$

for any small enough $\varepsilon > 0$. Hence, for large samplings only the critical points with small distance values are relevant for sketching $\mathrm{supp}(\mu)$.

Theorem 5. *Given a sequence of sample points $x_1,\ldots,x_n$ drawn independently from a probability measure μ with compact support on $\mathbb{R}^d$. If the reach ρ of the support is positive, then, for every $0<\varepsilon<\rho/2$,*

$$\lim_{n\to\infty} P\big[\mathit{compl}_\varepsilon(\mu)\ \mathit{contains\ a\ critical\ point\ of}\ d_n\big] = 0,$$

where $d_n:\mathbb{R}^d\to\mathbb{R}$ is the distance function to the set $X=\{x_1,\ldots,x_n\}$.

Proof. Let (x_n) be a sequence of points in $\mathrm{supp}(\mu)$ such that $c_n\in\mathrm{compl}_\varepsilon(\mu)$ is a critical point of d_n, i.e., the distance function to the first n points of the sequence. Since the closure of $\mathrm{compl}_\varepsilon(\mu)$ is compact we can assume by turning to an appropriate subsequence that the sequence (c_n) converges to $c\in\mathrm{compl}_\varepsilon(\mu)$. By the same argument we can even assume that all the c_n have the same index $i\in\{1,\ldots,d\}$. Let $y_{0n},\ldots,y_{in}$ be the points in $N(c_n)\subset X$ such that c_n is the center of the smallest enclosing ball of $\{y_{0n},\ldots,y_{in}\}$, i.e., this ball is given as $B(c_n,\|c_n-y_0\|)$ and does not contain any point from X in its interior. By the compactness of $\mathrm{supp}(\mu)$ we can assume that the sequence (y_{jn}) converges to $y_j\in\mathrm{supp}(\mu)$. Since c_n is the

center of the smallest enclosing ball of the points $y_{0n}, \ldots, y_{in}$ it can be written as a convex combination of these points, i.e.,

$$c_n = \sum_{j=0}^{i} \lambda_{jn} y_{jn}$$

with

$$\sum_{j=0}^{i} \lambda_{jn} = 1 \text{ and } \lambda_{jn} \geq 0,\ j = 0, \ldots, i.$$

That is, the vector $\lambda_n = (\lambda_{0n}, \ldots, \lambda_{in})$ is from the i-dimensional standard simplex which is compact. Hence by turning to yet another subsequence we can assume that λ_n converges in the standard simplex. Let $\lambda = (\lambda_0, \ldots, \lambda_i)$ be the limit of (λ_n), then we have $c = \sum_{j=0}^{i} \lambda_j y_j$ and thus c is the center of the smallest enclosing ball $B(c, \|c - y_0\|)$ of the points $y_0, \ldots, y_i$. If $B(c, \|c - y_0\|)$ does not contain any point from $\operatorname{supp}(\mu)$ in its interior, then c must be a point of the medial axis $\operatorname{ma}(\mu)$ which is impossible since the points $c_n \in \operatorname{compl}_\varepsilon(\mu)$ are at distance at least ε from the medial axis, and hence (c_n) can not converge to c. Thus, $B(c, \|c - y_0\|)$ must contain a point $z \in \operatorname{supp}(\mu)$ in its interior, i.e., there exists $\delta > 0$ such that $B(z, \delta) \subset B(c, \|c - y_0\|)$. Since c_n converges to c and the radii $\|c_n - y_{0n}\|$ converge to the radius $\|c_n - y_0\|$ we also have $B(z, \delta) \subset B(c_n, \|c_n - y_{0n}\|)$ for n large enough, and thus $\lim_{n\to\infty} \|x_1(z) - z\| \geq \delta$. That is, for n large enough the event

$$\big[\operatorname{compl}_\varepsilon(\mu) \text{ contains a critical point of } d_n\big]$$

implies the event $\big[\|x_1(z) - z\| \geq \delta\big]$. Hence,

$$\lim_{n\to\infty} P\big[\operatorname{compl}_\varepsilon(\mu) \text{ contains a critical point of } d_n\big] > 0.$$

implies that

$$\lim_{n\to\infty} P\big[\|x_1(z) - z\| \geq \delta\big] > 0 \quad \text{for} \quad z \in \operatorname{supp}(\mu),$$

which contradicts Lemma 2. Thus we have

$$\lim_{n\to\infty} P\big[\operatorname{compl}_\varepsilon(\mu) \text{ contains a critical point of } d_n\big] = 0.$$

□

In practice we expect that the number of critical points whose distance value is at most $\alpha \geq 0$ is increasing fast with growing α for small values of α. Once α is large enough such that all the critical points that belong to $\operatorname{supp}(\mu)$ have been found, the number of critical points remains constant for growing α, and is only increasing again once the critical points that belong to $\operatorname{ma}(\mu)$ are being discovered. There are two things one should bear in mind though. First, this behavior is only expected if $\operatorname{supp}(\mu)$ does not exhibit geometric features on different scales, because otherwise critical points that belong to the medial axis can be discovered before critical points that belong to the support, and second, by construction the medial axis $\operatorname{ma}(\mu)$ is sampled much more sparsely by critical points than $\operatorname{supp}(\mu)$. Hence, if $\operatorname{supp}(\mu)$ does not exhibit geometric features on different scales, then we expect the number of critical points to grow at first with growing α and to remain almost constant once all the critical points that belong to $\operatorname{supp}(\mu)$ have been discovered. A good value for α should be the point at which the number of critical points stays almost constant.

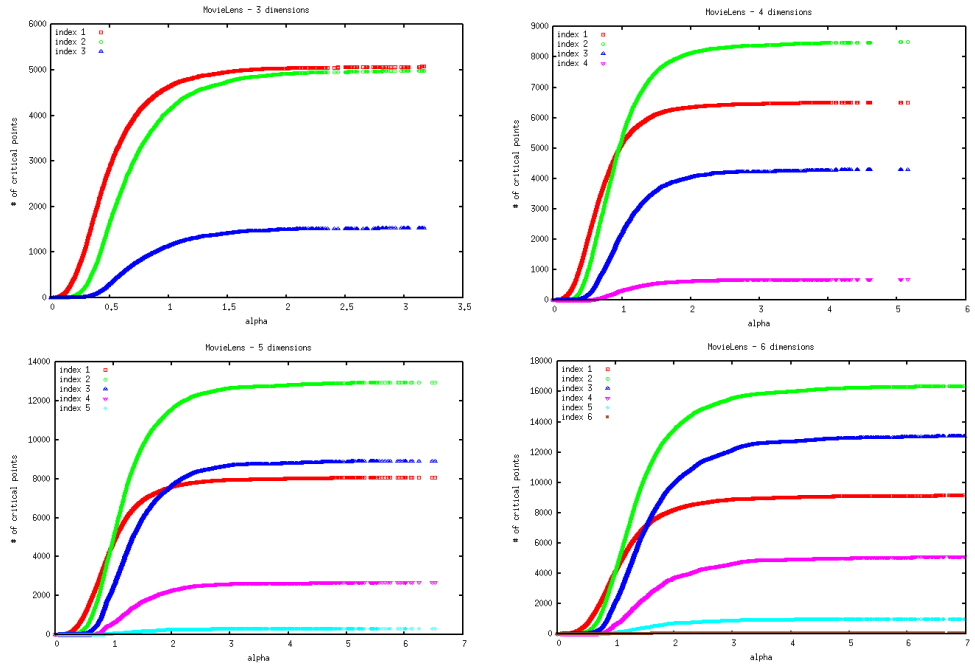

Figure 4: The number of critical points of the α-flow complex as a function of α for the first three to six dimensions (from top-left to bottom-right) for the MovieLens data set. Note that in dimension d we can only have critical points of index up to d.

6 Implementation

We have designed and implemented an algorithm for computing the Hasse diagram of the whole flow complex. The experimental results that we report in the following section have been obtained using this implementation.

There is a straightforward way to distribute our algorithm if we are only interested in computing the α-flow complex for small values of $\alpha > 0$, e.g., when sketching the support $\mathrm{supp}(\mu)$ of a probability measure μ on $\mathbb{R}^d$ from a given finite sampling $\{x_1, \ldots, x_n\}$ drawn independently from μ. The idea for distributing the algorithm is based on the following simple observation which is implied by the triangle inequality.

Observation 6. *For any $\alpha > 0$, if c is a critical point of the distance function $d_n : \mathbb{R}^d \to \mathbb{R}$ to the set $\{x_1, \ldots, x_n\}$ whose distance value $d_n(c)$ is at most α and whose nearest neighbor set $N(c)$ contains x_i, then $N(c)$ is contained in the ball $B(x_i, 2\alpha)$, i.e., $N(c) \subset B(x_i, 2\alpha)$.* □

The distributed algorithm can now be implemented through the following map- and reduce steps.

Map. For every $x_i \in X = \{x_1, \ldots, x_n\}$ let $X_i = B(x_i, 2\alpha) \cap \{x_1, \ldots, x_n\}$. For $i = 1, \ldots, n$, compute the α-flow complex for X_i. This can be done by computing the whole flow complex, i.e., the ∞-flow complex, for X_i and removing all critical points with distance value larger than α.

Reduce. Let $G = (V, E)$ be the graph whose vertex set is $V = [n] = \{1, \ldots, n\}$ and whose edge set is $E = \big\{(i, j) \in [n] \times [n] \,|\, X_i \cap X_j \neq \emptyset\big\}$. Combine the α-flow complexes for the sets X_i by traversing the connected components of the graph G in a breadth-first manner. Note that the α-flow complex is itself a graph, namely a

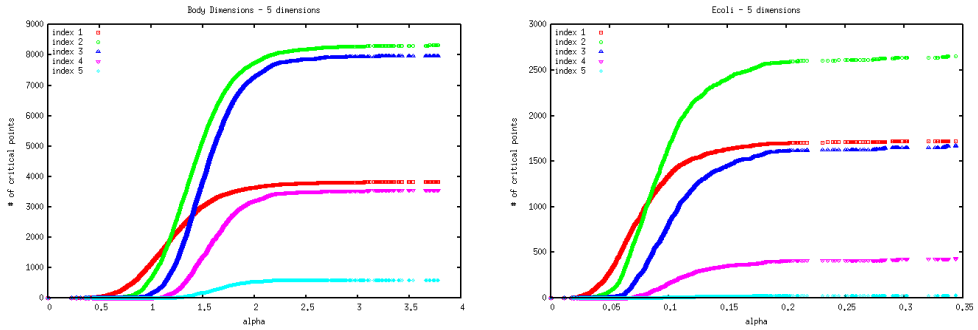

Figure 5: The number of critical points of the α-flow complex as a function of α for the Body Dimensions data set restricted to the first five dimensions (on the left) and the Ecoli data set (on the right).

Hasse diagram. The combination of two α-flow complexes is achieved by identifying all common vertices in the respective Hasse diagrams.

Theorem 7. *The distributed algorithm that comprises the map- and reduce step computes the α-flow complex of $\{x_1, \ldots, x_n\}$.*

Proof. We need to argue that the algorithm finds all critical points of the distance function d_n and connects them in the right way. By Observation 6 the α-flow complex of X_i does contain any critical point c of the distance function d_n with $x_i \in N(c)$ whose distance function value is at most α. Hence, any critical point of d_n with distance value at most α is contained in the union of the α-flow complexes of the sets X_i, where they are also connected in the right way. □

7 Experiments

We have tested our approach on three publicly available data sets in medium dimensions that we describe in the following.

MovieLens. The MovieLens 100k data set [21] was collected by the GroupLens Research Project at the University of Minnesota. It consists of 100,000 ratings from 943 users on 1,682 movies. The data set can be viewed as an incomplete matrix that is indexed by the users and the movies, respectively, where the matrix entries are the ratings. Therefore, the MovieLens data set is not a point cloud data set itself, but it is straightforward to derive point clouds for the movies and for the users, respectively, from the completed ratings matrix using principal component analysis. We used a technique by Bell et al. [2] (called ComputeNextFactor) for completing the ratings matrix that at the same time computes a low dimensional spectral embedding for the movies, i.e., every point in the low dimensional point cloud corresponds to a movie. Using this technique we created a 14-dimensional embedding of the 1682 movies. Figure 3 shows three rows out of a scatter plot matrix visualization of the data set, with scatter plots of the first three dimensions against all 14 dimensions. As can be seen, the trailing dimensions are correlated with the leading three dimensions and thus contribute less geometric information than the leading dimensions. We therefore computed α-flow complexes only for the data sets in three to six dimensions. Figure 4 shows the number of critical points as a function of the value α. The functions look like expected, namely we observe a fast increase in the number of critical points up to a threshold value for α. Beyond the threshold value the number of critical points stays almost constant. Note that the

threshold value increases with the dimension from ≈ 2 in three dimensions to ≈ 3 in six dimensions. This increase is expected since the distances between the points that represent the movies also increase with the dimension. Another interesting observation is the following: the plots in Figure 4 for Dimensions 5 and 6 indicate that the intrinsic dimension of the data set is four since almost no critical points of index six and only very few critical points of index five can be found.

Body Dimensions. The Body Dimensions data set [15] contains 507 points with 21 attributes (excluding four nominal attributes) that represent measurements of the human body. The first nine attributes are skeletal measurements, whereas the latter 12 are girth measurements. The first five skeletal measurements regard the body's torso. Here we restricted ourselves to these first five dimensions. Figure 5 (on the left) shows the number of critical points as a function of the value α for this data set. The function again looks like expected. We observe a fast increase in the number of critical points up to a threshold value for α which is ≈ 2.

Ecoli. The Ecoli data set [18] contains 336 points in eight dimensions. From these dimensions we removed two binary attributes and the sequence number and considered only the remaining five metric (Euclidean) dimensions. Figure 5 (on the right) shows the number of critical points as a function of the value α for this data set. Again, this function looks like expected. The threshold value for α here is ≈ 1.75.

8 Conclusions

We have presented an approach to sketch the compact support of a probability measure on $\mathbb{R}^d$ by an α-flow complex. With high probability, the α-flow complex is homotopy equivalent to the support of the measure for large enough samplings and good values for α. We have shown how to choose a good value for α in theory and in practice (on some real data sets). We have also briefly discussed a distributed algorithm to compute α-flow complexes for small values of α.

References

[1] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

[2] Robert Bell, Yehuda Koren, and Chris Volinsky. Modeling relationships at multiple scales to improve accuracy of large recommender systems. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 95–104. ACM, 2007.

[3] Jean-Daniel Boissonnat and Arijit Ghosh. Manifold reconstruction using tangential Delaunay complexes. In *Proceedings of the ACM Symposium on Computational Geometry (SOCG)*, pages 324–333, 2010.

[4] Frédéric Cazals and David Cohen-Steiner. Reconstructing 3D compact sets. *Computational Geometry: Theory and Applications*, 45(1-2):1–13, 2012.

[5] Frédéric Chazal, David Cohen-Steiner, and André Lieutier. A sampling theory for compact sets in Euclidean space. In *Symposium on Computational Geometry (SOCG)*, pages 319–326, 2006.

[6] Luc Devroye, Laszlo Györfi, and Gabor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1991.

[7] Tamal K. Dey, Joachim Giesen, and Matthias John. Alpha-shapes and flow shapes are homotopy equivalent. In *Annual ACM Symposium on Theory of Computing (STOC)*, pages 493–502, 2003.

[8] Tamal K. Dey, Joachim Giesen, Edgar A. Ramos, and Bardia Sadri. Critical points of the distance to an epsilon-sampling of a surface and flow-complex-based surface reconstruction. In *Symposium on Computational Geometry (SOCG)*, pages 218–227, 2005.

[9] Herbert Edelsbrunner. Surface Reconstruction by Wrapping Finite Sets in Space. In *The Goodman-Pollack Festschrift (Algorithms and Combinatorics)*, pages 379–404. Springer, 2003.

[10] Herbert Edelsbrunner, Michael Facello, and Jie Liang. On the Definition and the Construction of Pockets in Macromolecules. *Discrete Applied Mathematics*, 88:83–102, 1998.

[11] Herbert Edelsbrunner and John Harer. *Computational Topology - an Introduction*. American Mathematical Society, 2010.

[12] Joachim Giesen and Matthias John. The flow complex: a data structure for geometric modeling. In *Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 285–294, 2003.

[13] Joachim Giesen, Edgar A. Ramos, and Bardia Sadri. Medial axis approximation and unstable flow complex. In *Symposium on Computational Geometry*, pages 327–336, 2006.

[14] K. Grove. Critical point theory for distance functions. In *Differential geometry: Riemannian geometry*, volume 54 of *Proc. Sympos. Pure Math.*, pages 357–385. Amer. Math. Soc., 1993.

[15] Grete Heinz, Louis J. Peterson, Roger W. Johnson, and Carter J. Kerk. Exploring relationships in body dimensions. *Journal of Statistics Education*, 11(2), 2003. Data set available at `http://www.amstat.org/publications/jse/v11n2/datasets.heinz.html`.

[16] André Lieutier. Any open bounded subset of r^n has the same homotopy type as its medial axis. *Computer-Aided Design*, 36(11):1029–1046, 2004.

[17] Stefano Melacci and Mikhail Belkin. Laplacian support vector machines trained in the primal. *Journal of Machine Learning Research*, 12:1149–1184, 2011.

[18] Kenat Nakai. Ecoli data set, 1996. Data set available at `http://archive.ics.uci.edu/ml/datasets/Ecoli`.

[19] Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the Homology of Submanifolds with High Confidence from Random Samples. *Discrete & Computational Geometry*, 39(1-3):419–441, 2008.

[20] Partha Niyogi, Stephen Smale, and Shmuel Weinberger. A Topological View of Unsupervised Learning from Noisy Data. *SIAM Journal on Computing*, 40(3):646–663, 2011.

[21] GroupLens research group University of Minnesota. Movielens data set, 2011. Data set available at `http://www.grouplens.org/node/73`.

[22] Dirk Siersma. Voronoi diagrams and morse theory of the distance function. In *Geometry in Present Day Science*, pages 187–208. World Scientific, 1999.