Foundations of Geometric Methods in Data Science

Jean-Daniel Boissonnat
Mathieu Carrière
Frédéric Cazals
INRIA

Master Science des Données et Intelligence Artificielle
2023

Lecture 2 : Manifold Learning

Towards a sampling theory for geometric objects

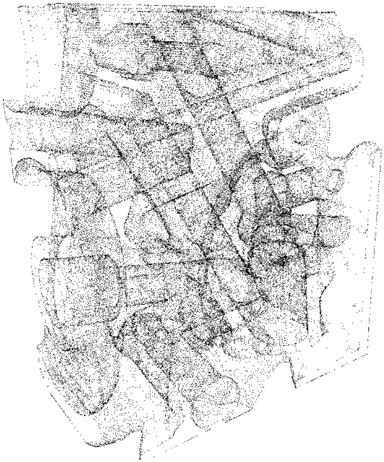
Topological and geometric models

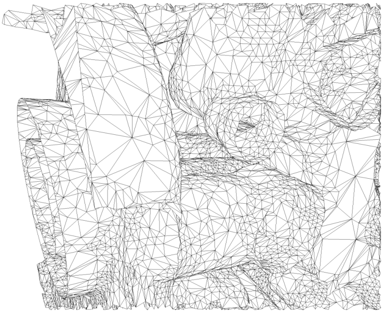
Distance functions and homotopic reconstruction

Interlude : molecules and affine diagrams

Reconstruction of submanifolds of $\mathbb{R}^d$

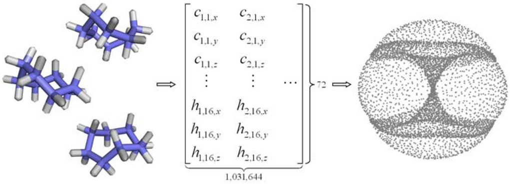
Reconstructing surfaces from point clouds





One can reconstruct a surface from 10^6 points within 1mn [CGAL]

Geometric data analysis

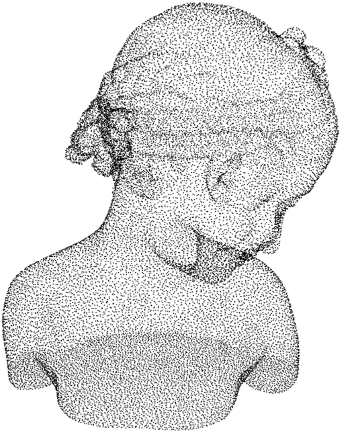


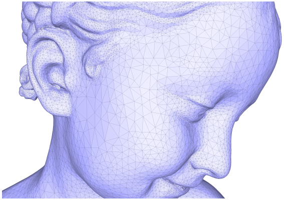
Geometrisation : Data = points + distances between points

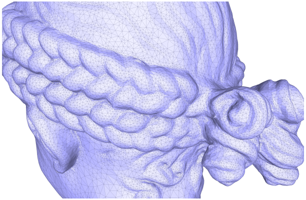
Manifold Hypothesis : Data lie close to a structure of “small” intrinsic dimension

Problem : Infer the structure from the data

Towards a sampling theory for geometric objects







- What spaces ?
- Quality criteria

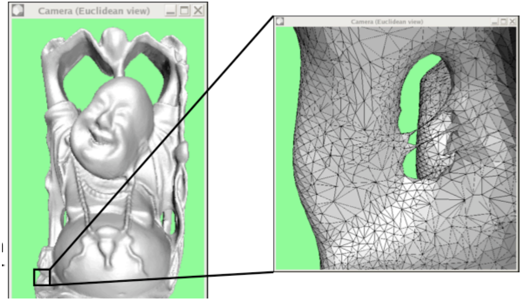
- Sampling conditions
- Reconstruction algorithms

Topological equivalence

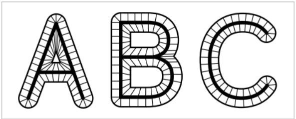
Homeomorphism

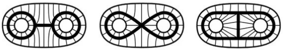
$f : X \rightarrow Y$ is a continuous bijective mapping whose inverse is continuous

$$X \approx Y$$



Deformation retraction



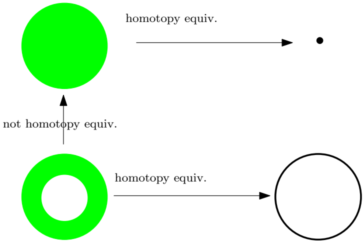


$$f_t : X \rightarrow A \subseteq X, \quad t \in [0,1] \quad \text{s.t.}$$

1. $(x, t) \rightarrow f_t(x)$ is continuous
2. $f_0(X) = id$
3. $f_1(X) = A$
4. $f_t|A = A$ for all t

A special case of a homotopy

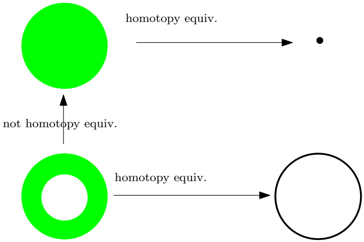
Homotopy equivalence



Intuitively, two spaces X and Y are homotopy equivalent if they can be transformed into one another

- by bending, shrinking and expanding operations
- but not by cutting or tearing

Homotopy equivalence



Intuitively, two spaces X and Y are homotopy equivalent if they can be transformed into one another

by bending, shrinking and expanding operations

but not by cutting or tearing

X is contractible if it is homotopy equivalent to a point

Homotopy and homotopy equivalence

Homotopy : a family of functions $f_t : X \rightarrow Y$ $t \in [0,1]$ s.t. $(x,t) \rightarrow f_t(x)$ is continuous

Two maps $f, g : X \rightarrow Y$ are homotopic, noted $f \simeq g$,

if there exists a homotopy joining them

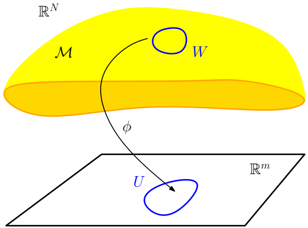
Two spaces X, Y are homotopy equivalent, noted $X \simeq Y$

if there exists maps $f : X \rightarrow Y$ and $g : Y \rightarrow X$ s.t. $fg \simeq id$ and $gf \simeq id$

If X deformation retracts onto Y, then $X \simeq Y$

Submanifolds of $\mathbb{R}^d$

A submanifold of dimension k is a subset of $\mathbb{R}^d$ that looks locally like (is homeomorphic to) an open set of an affine space of dimension k



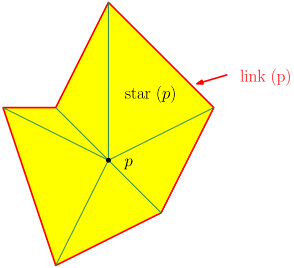
A curve a 1-dimensional submanifold

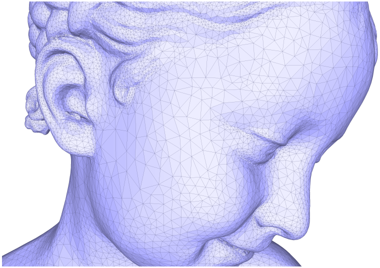
A surface is a 2-dimensional submanifold

Combinatorial (PL) manifolds

Definition

A pure simplicial complex $\hat{S}$ is a PL manifold of dimension k iff the link of each vertex is PL-homeomorphic to of a topological sphere of dimension $k-1$

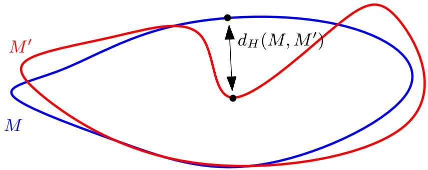




The underlying space of a PL manifold is a topological manifold

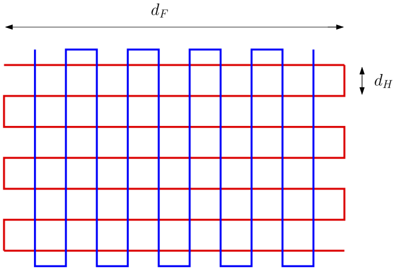
Geometric approximation of shapes

1. Hausdorff distance / Fréchet distance



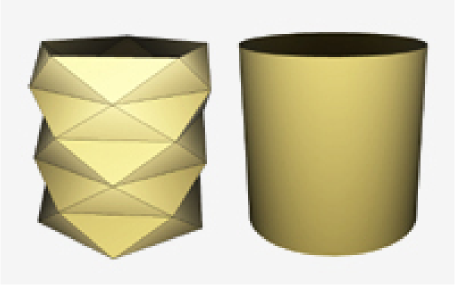
$$d_H(M, M') = \max\left(\sup_{x\in M}\inf_{x'\in M'}\|x - x'\|,\ \ \sup_{x\in M}\inf_{x'\in M'}\|x - x'\|\right)$$

$$= \inf\{r : M \subset M'^{+r} \text{ and } M' \subset M^{+r}\}$$



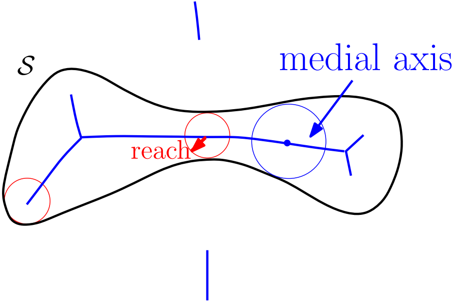
Geometric approximation of shapes

2. Tangent spaces approximation



Reach

Captures curvature and bottlenecks



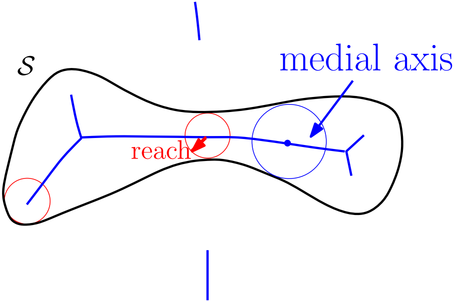
H. Federer

Local feature size

$\forall x \in \mathcal{S}, \quad \mathrm{lfs}(x) = d(x, \mathrm{axis}(\mathcal{S}))$ (1-Lipschitz : $|f(x) - f(y)| \leq \|x - y\|$)

$\mathrm{rch}(\mathcal{S}) = \inf_{x \in \mathcal{S}} \mathrm{lfs}(x)$

Sampling conditions and ε-nets

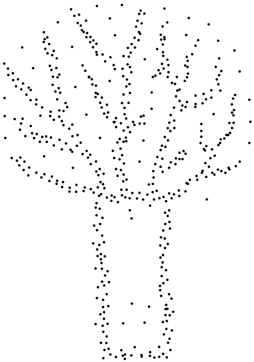


$(\epsilon, \bar{\eta})$-net of $\mathcal{S}$

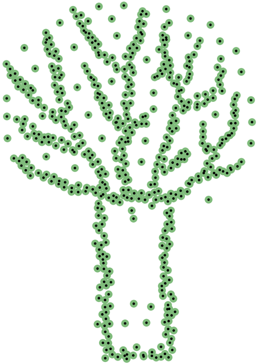
1. Covering: $\mathcal{P} \subset \mathcal{S}, \forall x \in \mathcal{S}, \quad d(x, \mathcal{P}) \leq \epsilon \, \mathrm{lfs}(x)$
2. Packing: $\forall p, q \in \mathcal{P}, \quad \|p - q\| \geq \bar{\eta}\varepsilon \, \min(\mathrm{lfs}(p), \mathrm{lfs}(q))$

Reconstruction of geometric shapes

Union of balls and distance functions



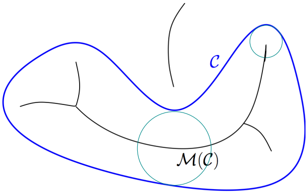
Sample P



Union of balls $P^{+\alpha}$

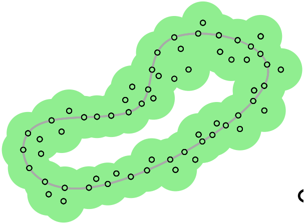
Reconstruction theorems

Union of balls and distance functions



Niyogi, Smale, Weinberger [2008]

If $\mathbb{M}$ is a submanifold of positive reach τ ,
P an ε-dense sample of $\mathbb{M}$,
then, for all $\alpha \in [\sim \varepsilon, \sim \tau]$, $P^{+\alpha} \simeq \mathbb{M}$



Chazal, Cohen-Steiner, Lieutier [2009]

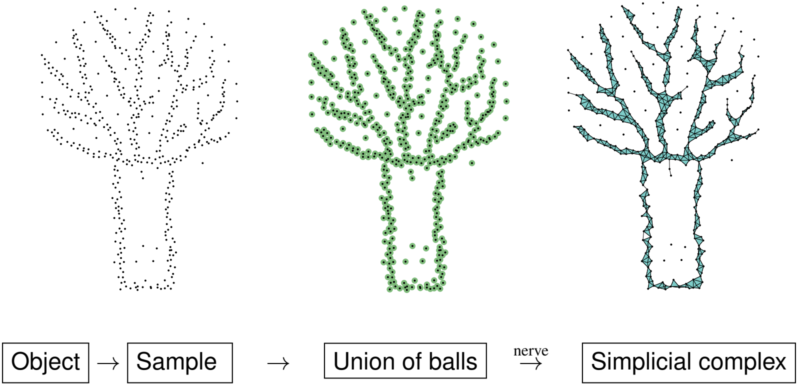
Extension to general compact sets

Chazal, Cohen-Steiner, Mérigot [2011]

Extension to points sets with outliers

Shape reconstruction

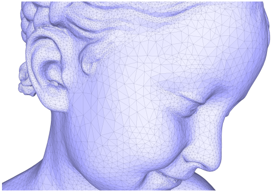
Discrete approximation of continuous spaces



Two issues

Curse of dimensionality: The Čech and the alpha-complex are big ($O(n^d)$ and $O(n^{d/2})$) and difficult to compute in high dimensions

Quality of approximation : Both complexes are not (in general) homeomorphic to X

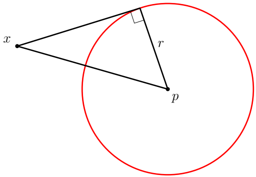


The manifold hypothesis: In many applications, the intrinsic dimension k is much smaller than the dimension d of the ambient space

- Can we bound the combinatorial complexity as a function of the intrinsic dimension ?
- Can we reconstruct a simplicial complex homeomorphic to the manifold, i.e. a triangulation of the manifold?

Power of a point wrt to a ball

Power of x wrt b : $D(x,b) = (x-p)^2 - r^2$



$$
\begin{aligned}
x \in \mathrm{int} b &\iff D(x,b) < 0 \\
x \in \partial b &\iff D(x,b) = 0 \\
x \notin b &\iff D(x,b) > 0
\end{aligned}
$$

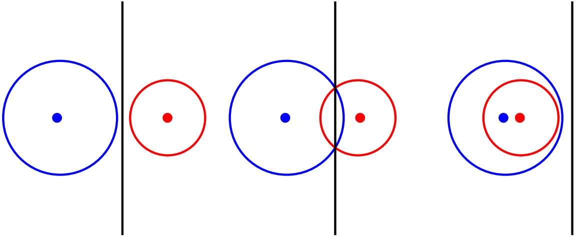
Remarks

- D is not a true distance
- We can consider r^2 as the weight of p and don't require it to be > 0

Radical hyperplane

- The set of points that have a same power wrt two balls $b_1(p_1, r_1)$ and $b_2(p_2, r_2)$ is a hyperplane

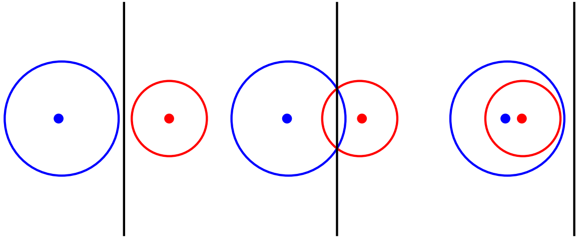
$$
\begin{aligned}
D(x, b_1) = D(x, b_2) &\iff (x - p_1)^2 - r_1^2 = (x - p_2)^2 - r_2^2 \stackrel{\text{def}}{=} r_x^2 \\
&\iff -2p_1 x + p_1^2 - r_1^2 = -2p_2 x + p_2^2 - r_2^2 \\
&\iff 2(p_2 - p_1)x + (p_1^2 - r_1^2) - (p_2^2 - r_2^2) = 0
\end{aligned}
$$



Radical hyperplane

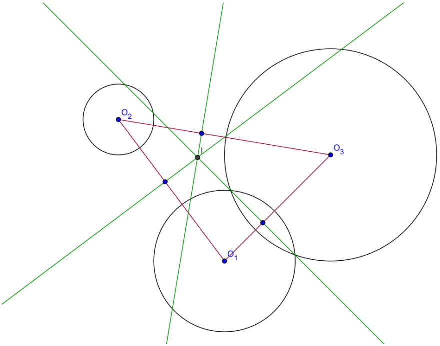
- The set of points that have a same power wrt two balls $b_1(p_1, r_1)$ and $b_2(p_2, r_2)$ is a hyperplane

$$
\begin{aligned}
D(x, b_1) = D(x, b_2) \quad &\iff \quad (x - p_1)^2 - r_1^2 = (x - p_2)^2 - r_2^2 \stackrel{\text{def}}{=} r_x^2 \\
&\iff \quad -2p_1 x + p_1^2 - r_1^2 = -2p_2 x + p_2^2 - r_2^2 \\
&\iff \quad 2(p_2 - p_1)x + (p_1^2 - r_1^2) - (p_2^2 - r_2^2) = 0
\end{aligned}
$$



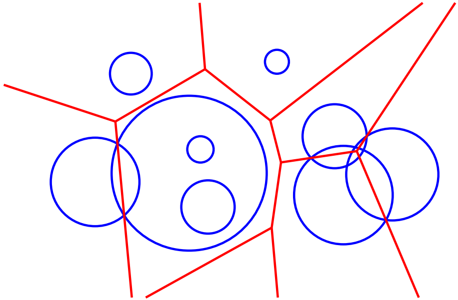
- The radical hyperplane is the set of centres x of the balls $B(x, r_x)$ that are orthogonal to b_1 and b_2

Radical centre



Voronoi diagrams of balls (or weighted points)

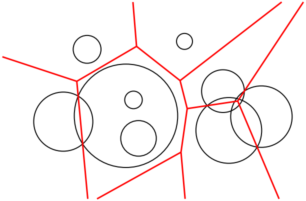
$B = \{b_1, ..., b_n\}$ $\qquad$ $D(x, b) = (x - p)^2 - r^2$



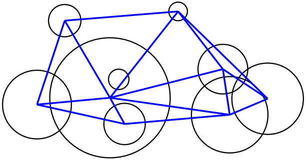
Voronoi cell : $V(b_i) = \{x : D(x, b_i) \leq D(x, b_j) \forall j\}$

Voronoi diagram of B : $\mathrm{Vor}(B) = \{$ set of cells $V(b_i), \quad b_i \in B\}$

Delaunay triangulations of balls (or weighted points)



$\mathrm{Vor}(B)$



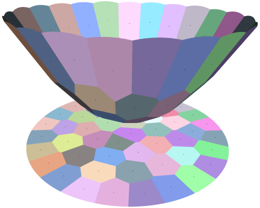
$\mathrm{Del}(B)$ is the nerve of $\mathrm{Vor}(B)$

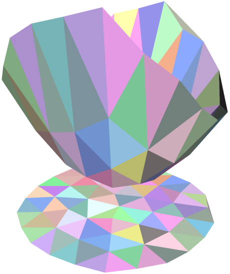
Theorem

If the balls are in general position, then $\mathrm{Del}(B)$ is a triangulation of a subset $\mathcal{P}' \subseteq \mathcal{P}$ of the points

Correspondence between structures

$h_{b_i} : x_{d+1} = 2p_i \cdot x - p_i^2 + r_i^2$ $\qquad$ $\hat{b}_i = (p_i, p_i^2 - r_i^2) = h_{b_i}^*$





$$\mathcal{V}(B) = h_{b_1}^+ \cap \ldots \cap h_{b_n}^+ \quad \xrightarrow{\text{duality}} \quad \mathcal{D}(B) = \text{conv}^-(\{\hat{b}_1, \ldots, \hat{b}_n\})$$

$\uparrow$ $\qquad\qquad\qquad\qquad$ $\downarrow$

Voronoi diagram of B $\xrightarrow{\text{nerve}}$ Delaunay triang. of B

The diagram commutes if B is in general position

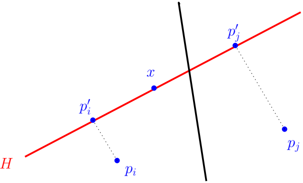
Affine diagrams

Sites + distance functions s.t. the bisectors are hyperplanes

Theorem [Aurenhammer]

Any affine diagram of $\mathbb{R}^d$ is the Voronoi diagram of a set of balls of $\mathbb{R}^d$

Intersection of a Voronoi diagram with a k-flat H of $\mathbb{R}^d$



$$\|x - p_i\|^2 \leq \|x - p_j\|^2$$

$$\Leftrightarrow \quad \|x - p'_i\|^2 - \|p_i - p'_i\|^2 \leq \|x - p'_i\|^2 - \|p_j - p'_j\|^2$$

Let $B = \{b_i = (p'_i, -\|p_i - p'_i\|^2)\}$ (weighted points in H)

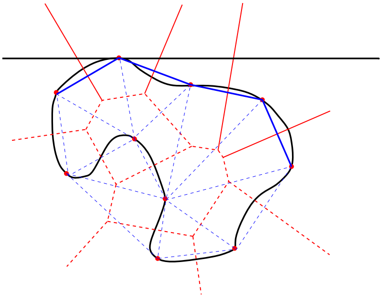
- $\text{Vor}(\mathcal{P}) \cap H \quad = \quad \text{Vor}(B)$ (a weighted Voronoi diagram in H)
- Can be computed in time $O(n^{\lfloor \frac{k+1}{2} \rfloor})$

(while the full diagram has complexity $\Theta(n^{\lfloor \frac{d+1}{2} \rfloor})$)

Restriction of a Delaunay triangulation to H

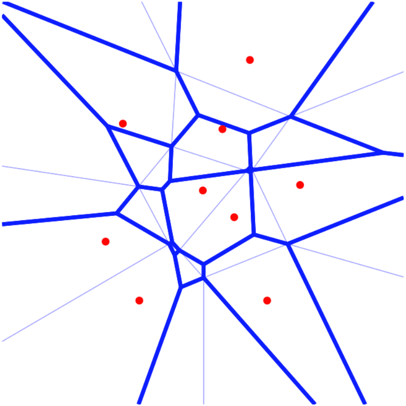
Definition: $\mathrm{Del}_{|H}(\mathcal{P})$ is the nerve of $\mathrm{Vor}(P) \cap H$

Equivalently, $\mathrm{Del}_{|H}(\mathcal{P})$ is the subcomplex of $\mathrm{Del}(\mathcal{P})$ consisting of the simplices that can be circumscribed by an empty ball centered on H



$$\mathrm{Del}_{|H}(\mathcal{P}) \stackrel{1-1}{\longleftrightarrow} \mathrm{Del}(B)$$

Voronoi diagram of order *k*



Each cell is the set of points that have the same *k* nearest sites

Voronoi diagrams of order k are weighted Voronoi diagrams

$S_1, S_2, \ldots$ the subsets of k points of $\mathcal{P}$

$$
\begin{aligned}
\delta(x, S_i) &= \frac{1}{k} \sum_{p \in S_i} (x-p)^2 \\
&= x^2 - \frac{2}{k} \sum_{p \in S_i} p \cdot x + \frac{1}{k} \sum_{p \in S_i} p^2 \\
&= D(b_i, x)
\end{aligned}
$$

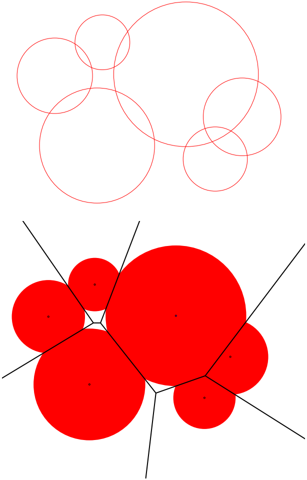
where b_i is the ball centered at $c_i = \frac{1}{k} \sum_{p \in S_i} p$

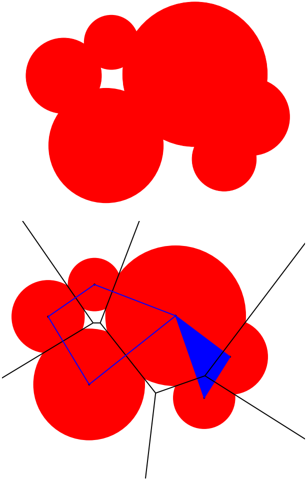
of radius $r_i^2 = c_i^2 - \frac{1}{k} \sum_{p \in S_i} p^2$

$$x \in \mathrm{Vor}_k(S_i) \quad \Leftrightarrow \quad \delta(x, S_i) \leq \delta(x, S_j) \ \ \forall j$$

Delaunay triangulation restricted to a molecule

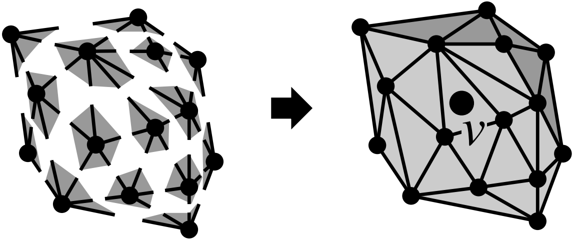
$U = \bigcup b_i,\ i = 1, \ldots, n$





$\mathrm{Del}_{|U}(B)$ is the nerve of the cover of U by the cells of $\mathrm{Vor}(B)$

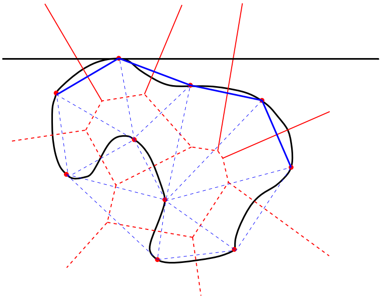
Triangulation of manifolds by star stitching

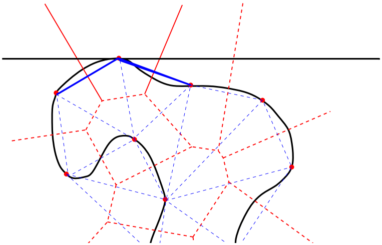


1. Construct local Delaunay triangulations (stars)
2. Insure that the local triangulations are consistent

   i.e. a simplex appears in the stars of all its vertices
3. Stitch the stars

The tangential Delaunay complex



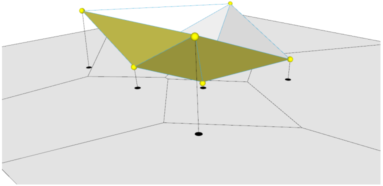


Local triangulations

$$\forall p \in \mathcal{P}: \qquad T_p(\mathcal{P}) = \mathrm{star}(p, \mathrm{Del}_{|T_p}))$$

Tangential complex

$$\mathrm{Del}_{T\mathbb{M}}(\mathcal{P}) = \{T_p(\mathcal{P}), p \in \mathcal{P}\}$$



Nice properties of the tangential Delaunay complex

- Subcomplex of $\mathrm{Del}(\mathcal{P})$:

  $\mathrm{Del}_{T\mathbb{M}}(\mathcal{P}) \subseteq \mathrm{Del}(\mathcal{P})$

  $\mathrm{Del}_{T\mathbb{M}}(\mathcal{P})$ is embedded in $\mathbb{R}^d$

- Dimension : The dimension of $\mathrm{Del}_{T\mathbb{M}}(\mathcal{P})$ is the dimension k of the submanifold $\mathbb{M}$ (under general position)

- Complexity :

  $\mathrm{Del}_{T\mathbb{M}}(\mathcal{P})$ can be computed without computing $\mathrm{Del}(\mathcal{P})$

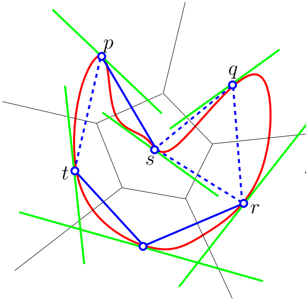
  If $\mathcal{P}$ is an ε-sample of $\mathbb{M}$, its complexity is $O(2^k|\mathcal{P}|)$ (linear in $|\mathcal{P}|$)
  and does not depend on d

Construction of $\mathrm{Del}_{T\mathbb{M}}(\mathcal{P})$

1. project $\mathcal{P}$ in T_p and weight the points accordingly $\rightarrow B_p$ (in time $O(dn)$)
2. construct $\mathrm{star}(p_i, \mathrm{Del}(B_p)) \subset T_{p_i}$ (in time $O(n^{\lfloor \frac{k+1}{2} \rfloor})$)
3. $\mathrm{star}(p_i, \mathrm{Del}_{T\mathbb{M}}(\mathcal{P})) \overset{1-1}{\longleftrightarrow} \mathrm{star}(p_i, \mathrm{Del}(B_p))$

Complexity : linear in d, exponential in k

Inconsistencies

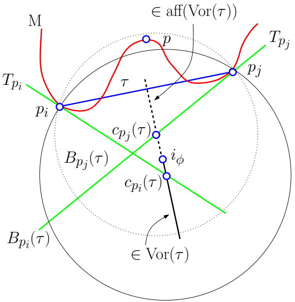


A simplex might not appear in the stars of all its vertices

$\Rightarrow \mathrm{Del}_{\mathcal{T}\mathbb{M}}(\mathcal{P})$ is not necessarily a PL manifold

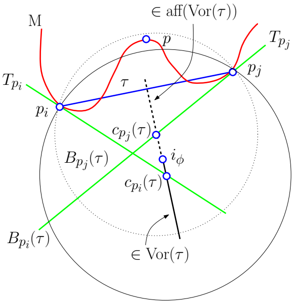
Inconsistency triggers

1. $\tau \in \mathrm{star}(p_i) \Rightarrow \quad B(c_{p_i}(\tau) \cap \mathcal{P} = \emptyset$
2. $\tau \notin \mathrm{star}(p_j) \Rightarrow \quad B(c_{p_j}(\tau) \cap \mathcal{P} = \mathcal{C} \neq \emptyset$
3. $\exists p \in \mathcal{C} : \quad \phi = \tau * p \in \mathrm{Del}(\mathcal{P})$
   $(\dim(\phi) = k + 1)$



Inconsistency triggers

1. $\tau \in \mathrm{star}(p_i) \Rightarrow \quad B(c_{p_i}(\tau) \cap \mathcal{P} = \emptyset$
2. $\tau \notin \mathrm{star}(p_j) \Rightarrow \quad B(c_{p_j}(\tau) \cap \mathcal{P} = \mathcal{C} \neq \emptyset$
3. $\exists p \in \mathcal{C} : \quad \phi = \tau * p \in \mathrm{Del}(\mathcal{P})$
   $(\dim(\phi) = k + 1)$



if the diameter of τ is small and thick

$\Rightarrow$ c_i et c_j are close & $\mathrm{aff}(\tau) \approx T_{p_i} \approx T_{p_j}$

$\Rightarrow$ $\exists$ a $(k+1)$-simplex ϕ which is not well "protected"

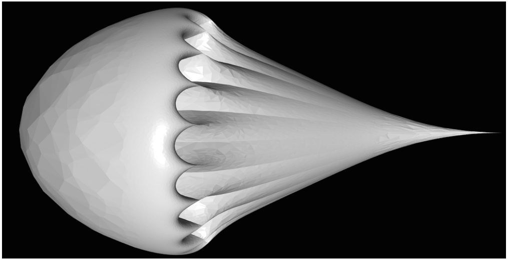
Such simplices can be removed by slightly perturbing the data

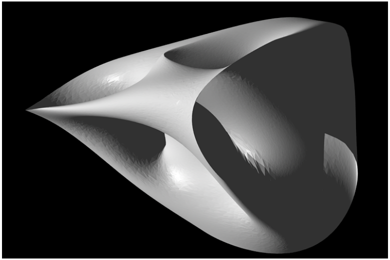
Further results

- Topological correctness
- Control on the Hausdorff distance
- Control on the angles between the simplices and the tangent spaces

Details in B., Chazal, Yvinec. Geometric and Topological Inference

Reconstruction of Rieman surfaces of $\mathbb{R}^8$





Data provided by A. Alvarez

Triangulation of the space of conformations of C_8H_{16}