Geometric clustering in the normed plane

Pedro Martín

University of Extremadura, Badajoz

Cáceres, March 2016

Geometric clustering

$\mathbb{M}^2 = (\mathbb{R}^2, \|\cdot\|)$ is a 2-dimensional *normed* (or *Minkowski*) *plane*.

Geometric clustering

$\mathbb{M}^2 = (\mathbb{R}^2, \|\cdot\|)$ is a 2-dimensional *normed* (or *Minkowski*) *plane*.
Let S be a set of n points in the normed plane and k a fixed number.

Geometric clustering

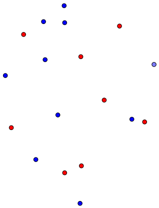
$\mathbb{M}^2 = (\mathbb{R}^2, \|\cdot\|)$ is a 2-dimensional *normed* (or *Minkowski*) *plane*. Let S be a set of n points in the normed plane and k a fixed number.

How can S be separated (by an algorithm) in k clusters verifying some conditions?

Geometric clustering

$\mathbb{M}^2 = (\mathbb{R}^2, \|\cdot\|)$ is a 2-dimensional *normed* (or *Minkowski*) *plane*. Let S be a set of n points in the normed plane and k a fixed number.

How can S be separated (by an algorithm) in k clusters verifying some conditions?



Geometric clustering

$k = 1$, minimizing the radius of a enclosing disc:

- Elzinga-Hearn and Shamos-Hoey (Euclidean plane).
- Alonso-Martini-Spirova and Jahn (general normed plane).

$k = 2$, minimizing the maximum Euclidean diameter of the clusters:

- Avis, $O(n^2 \log n)$.
- Asano-Bhattacharya-Keil-Yao, $O(n \log n)$.

$k = 2$, minimizing the sum of the two Euclidean diameters:

- Monma-Suri, $O(n^2)$.

$k = 2$, μ a measure, $\mu_1 > 0$ and $\mu_2 > 0$, splitting S into two clusters A and B such that $\mu(A) \leq \mu_1$ and $\mu(B) \leq \mu_2$:

- Hershberger and Suri,
  - $\mu =$Euclidean diameter, $O(n \log n)$.
  - $\mu =$area, perimeter, or diagonal of the smallest rectangle with sides parallel to the coordinates axes ($O(n \log n)$ time).
  - $\mu =$radius of the smallest enclosing sphere with the norms L_1 ($O(n \log n)$ time) or the Euclidean norm ($O(n^2 \log n)$ time)

Geometric clustering

$k = 2$, the 2-center problem: cover S by (the union of) two congruent closed disks whose radius is as small as possible.

- Eppstein and Sharir (1997), near linear time cost (Euclidean case).

$k = 3$, minimizing the maximum Euclidean diameter

- Hagauer-Rote, $O(n^2 \log^2 n)$

Any k, minimizing any monotone function $\mathcal{F}$ ($\mathcal{F} : \mathbb{R}^k \to \mathbb{R}$) of the Euclidean diameters or the Euclidean radii of the clusters.
Examples of $\mathcal{F}$:

- The sum of the diameters (or the radii)
- The maximum of the diameters (or the radii)
- The sum of the squares of the diameters (or the radii).

- Capoyleas-Rote-Woeginger, polynomial time.

Hagauer-Rote and Capoyleas-Rote-Woeginger obtain their results from this theorem

Theorem (Capoyleas-Rote-Woeginger)

Let A and B be two sets of points in the Euclidean plane. Then, there are two linearly separable sets A' and B' such that $\mathrm{diam}(A') \leq \mathrm{diam}(A)$, $\mathrm{diam}(B') \leq \mathrm{diam}(B)$, *and* $A' \cup B' = A \cup B$.

Figure: Non linearly separable (left) and linarly separable sets (right)

sets 2.pdf

Linear separation of clusters

This first statement is used in the proof of the Theorem:
In every triangle with an obtuse angle, the side lying opposite to the obtuse angle is the (Euclidean) longest side in the triangle.

Linear separation of clusters

This first statement is used in the proof of the Theorem:
In every triangle with an obtuse angle, the side lying opposite to the obtuse angle is the (Euclidean) longest side in the triangle.

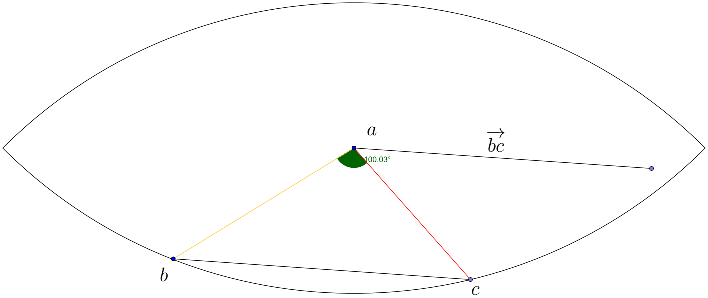


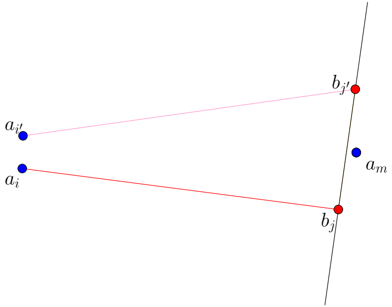
Figure: The side opposite to the obtuse angle is not the longest side in in the triangle $\triangle abc$.

Linear separation of clusters

This second statement is used in the proof of Theorem:

1.$\mathrm{diam}(A) \geq \mathrm{diam}(B)$
2.$\{a_i, a_i', a_m\} \subset A, \{b_j, b_j'\} \subset B$
Clockwise order: $a_{i'}, b_{j'}, a_m, b_j, a_i$
3. $< b_j, b_{j'} >$ separates $\{a_i, a_{i'}\}$ from a_m.

$\Longrightarrow_{(\mathbb{E}^2)}$ $\{\|a_i - b_j\|, \|a_{i'} - b_{j'}\|\} \leq \mathrm{diam}(A)$.



Linear separation of clusters

But this point configuration is possible in a general normed plane:

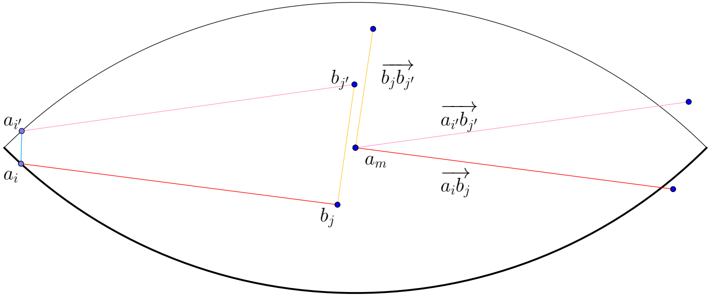
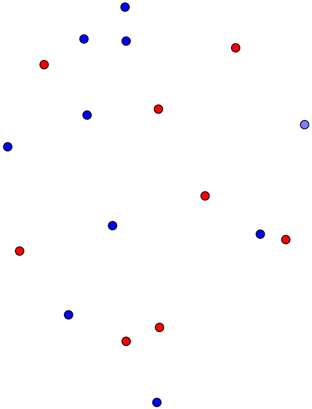


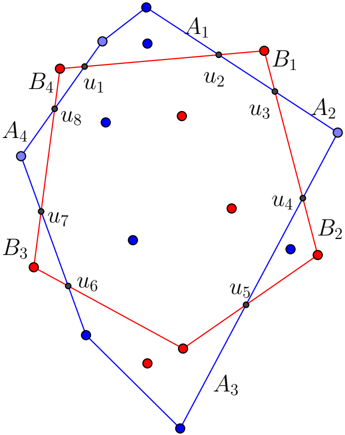
Figure: $\|a_i - b_j\|$ and $\|a_{i'} - b_{j'}\|$ are longer than the diameter of A.

Objective: to prove the Theorem for any normed plane.



Linear separation of clusters

Step 1: $\{u_1, u_2, \ldots, u_{2k}\} = \partial(\mathrm{conv}(A)) \cap \partial(\mathrm{conv}(B))$.



We can assume that $\mathrm{diam}(A) \geq \mathrm{diam}(B)$

We say that...

- (A_i, B_j) is a *bad pair* if $\mathrm{diam}(A_i \cup B_j) > \mathrm{diam}(A)$.

  Then, A_i and B_j are *bad partners*.

- $a_i \in A_i$ and $b_j \in B_j$ are *bad points* if $\|a_i - b_j\| > \mathrm{diam}(A)$.

  Then, a_i and b_j are *bad partners*,

  and the segment $\overline{a_i b_j}$ is a *bad segment*.

Lemma

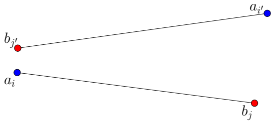
Let (A_i, B_j) and $(A_{i'}, B_{j'})$ two disjoint bad pairs. Let us choose $a_i \in A_i, b_j \in B_j, a_{i'} \in A_{i'}, b_{j'} \in B_{j'}$ such that $\overline{a_i b_j}$ and $\overline{a_{i'} b_{j'}}$ are bad segments. Then, either these bad segments intersect, or any point $a \in A_m$ belonging to the halfplane defined by $< b_j b_{j'} >$ where a_i and $a_{i'}$ are not contained, is not bad.

Lemma

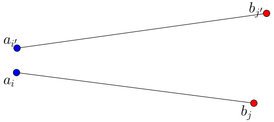
Let (A_i, B_j) and $(A_{i'}, B_{j'})$ two disjoint bad pairs. Let us choose $a_i \in A_i, b_j \in B_j, a_{i'} \in A_{i'}, b_{j'} \in B_{j'}$ such that $\overline{a_i b_j}$ and $\overline{a_{i'} b_{j'}}$ are bad segments. Then, either these bad segments intersect, or any point $a \in A_m$ belonging to the halfplane defined by $< b_j b_{j'} >$ where a_i and $a_{i'}$ are not contained, is not bad.

Skecth of the proof. Possible clockwise order (up to symmetries):

Case 1: $a_i, b_{j'}, a_{i'}, b_j$



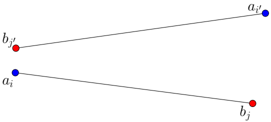
Case 2: $a_i, a_{i'}, b_{j'}, b_j$



Linear separation of clusters

Case 1: clockwise order

$a_i, b_{j'}, a_{i'}, b_j$



We get a contradiction:

$$\mathrm{diam}(A) + \mathrm{diam}(B) \geq \|a_i - a_{i'}\| + \|b_j - b_{j'}\| \geq \|a_i - b_j\| + \|a_{i'} - b_{j'}\| > 2\,\mathrm{diam}(A).$$

Linear separation of clusters

Case 2: clockwise order $a_i, a_{i'}, b_{j'}, b_j$:

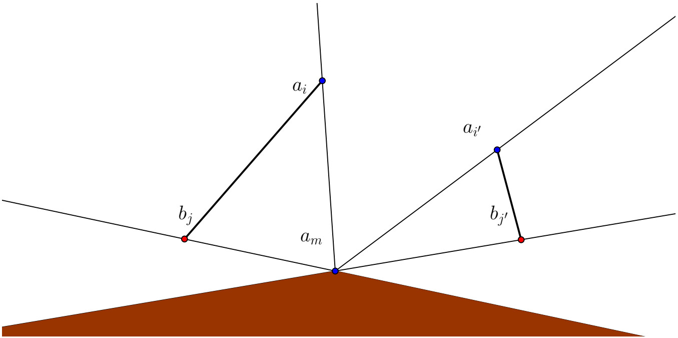


Figure: (a_i, b_j), $(a_{i'}, b_{j'})$ are bad partners $\Longrightarrow \nexists$ any bad partner for a_m

Linear separation of clusters

Case 2: clockwise order $a_i, a_{i'}, b_{j'}, b_j$:

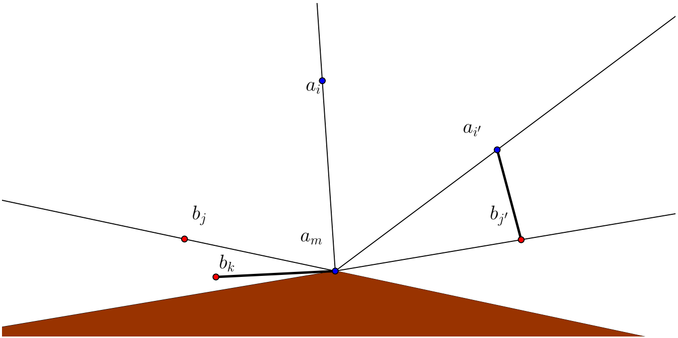
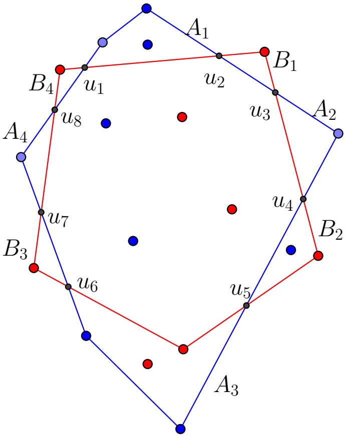


Figure: (a_i, b_j) and $(a_{i'}, b_{j'})$ bad partners $\Longrightarrow \nexists$ any bad partner for a_m

Step 2: Maximal cyclic subsequences of polygons.



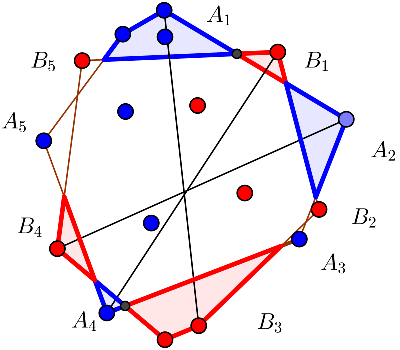
Linear separation of clusters

Step 2: Maximal cyclic subsequences of polygons.

- Consider maximal cyclic subsequences of adjacent bad polygons A_i.
  - No "good" polygon A_k belongs to one of this maximal cyclic subsequences of bad A_i-polygons.
  - Some intervening "good" polygon B_j can belong to this maximal cyclic subsequences of A_i-polygons.
- Similarly with adjacent bad polygons B_j.
- These maximal cyclic sequences are noted by $\bar{\mathbf{A}}_1, \bar{\mathbf{A}}_2, \ldots, \bar{\mathbf{A}}_p$ and $\bar{\mathbf{B}}_1, \bar{\mathbf{B}}_2, \ldots, \bar{\mathbf{B}}_q$.

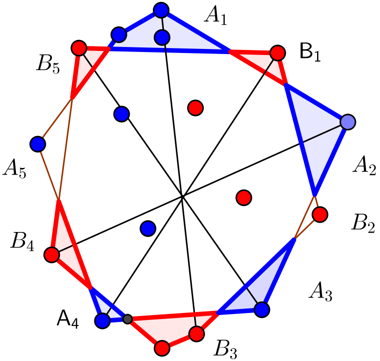
Example with 3 maximal cyclic subsequences of A_i-polygons and 3 maximal subsequences of B_j-polygons:

$\bar{\mathbf{A}}_1 = \{A_1\}$
$\bar{\mathbf{B}}_1 = \{B_1\}$
$\bar{\mathbf{A}}_2 = \{A_2\}$
$\bar{\mathbf{B}}_2 = \{B_3\}$
$\bar{\mathbf{A}}_3 = \{A_4\}$
$\bar{\mathbf{B}}_3 = \{B_5\}$



Example with 3 maximal cyclic subsequences of A_i-polygons, 3 maximal subsequences of B_j-polygons, and "good" intervening polygons:

$\bar{\mathbf{A}}_1 = \{A_1\}$
$\bar{\mathbf{B}}_1 = \{B_1\}$
$\bar{\mathbf{A}}_2 = \{A_2, B_2, A_3\}$
$\bar{\mathbf{B}}_2 = \{B_3\}$
$\bar{\mathbf{A}}_3 = \{A_4\}$
$\bar{\mathbf{B}}_3 = \{B_4, A_5, B_5\}$



Linear separation of clusters

Properties

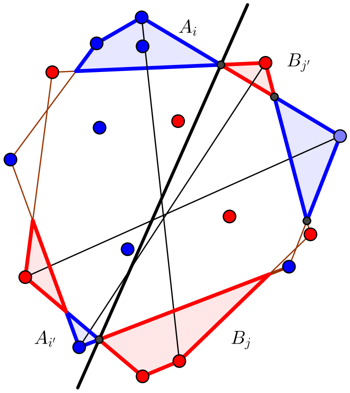
- Let (A_i, B_j) and $(A_{i'}, B_{j'})$ be two disjoint bad pairs. Then

$$A_i, A_{i'} \in \bar{\mathbf{A}}_k \Longrightarrow B_j, B_{j'} \in \bar{\mathbf{B}}_t$$

- The number of maximal cyclic sequences of adjacent bad A_i-polygons and B_j-polygons is the same.
- If $(\bar{\mathbf{A}}_i, \bar{\mathbf{B}}_j)$ and $(\bar{\mathbf{A}}_{i'}, \bar{\mathbf{B}}_{j'})$ are disjoint bad pairs of maximal subsequences, then there exist two (one from every pair) bad-crossing segments.
- There is an odd number of subsequences from each cluster, and they must be completely interlacing.

Linear separation of clusters

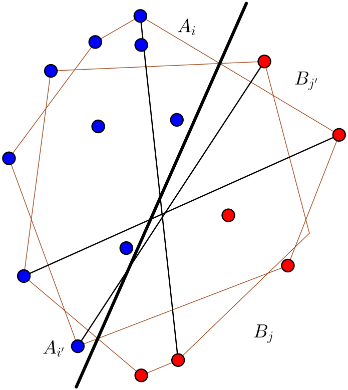
Step 3: Separate the sets.



Linear separation of clusters

- Let A_i be the last polygon of a maximal cyclic subsequence (in clockwise order)
- Let B_j be the last bad partner of A_i.
- Let $B_{j'}$ be the first bad polygon after A_i
- let $A_{i'}$ be the first bad partner of $B_{j'}$.
- Choose the line L going through the point just before B_j and the point just after $B_{j'}$.
- Define B' to be the points in $A \cup B$ lying on the same side of L as B_j and $B_{j'}$, and A' as the remaing points.

Linear separation of clusters



Linear separation of clusters

Proposition

$$\mathrm{diam}(A') \leq \mathrm{diam}(A), \qquad \mathrm{diam}(B') \leq \mathrm{diam}(B).$$

Theorem

Let A and B be two sets of points in a general normed plane. Then, there are two linearly separable sets A' and B' such that $\mathrm{diam}(A') \leq \mathrm{diam}(A)$, $\mathrm{diam}(B') \leq \mathrm{diam}(B)$, and $A' \cup B' = A \cup B$.

Corollary

In the construction in the Theorem,

$$\mathrm{perimeter}(A) + \mathrm{perimeter}(B) \geq \mathrm{perimeter}(A') + \mathrm{perimeter}(B')$$

holds. If $\mathrm{conv}(A) \cap \mathrm{conv}(B) \neq \emptyset$, then the inequality is strict.

Some consequences

The 2-clustering problem for diameter respect to the minimum: Dividing S in two sets minimizing the maximum diameter of the sets.

Theorem

Given a set S of n points in a normed plane, the 2-clustering problem for diameter respect to the minimum can be computed in $O(n^2 \log^2 n)$ time.

- Sort the distances d_i between the points of S into increasing order.
- By a binary search, locate the minimum d_i that admits a *stabbing line* for the set of segments meeting point of S at distance greater than d_i.

Some consequences

The k-clustering problem for diameter respect to a function $\mathcal{F}$ (for example, $\mathcal{F}$ can be the *maximum*, the *sum*, or the *sum of squares*):

Dividing S in k sets minimizing a function $\mathcal{F}$ of the diameters of the sets.

Theorem

Consider the optimal k-clustering problem for the diameter respect to a monotone increasing function $\mathcal{F}$ of such as diameters. For every set S of n points in a general normed plane,

- *There is an optimal k-clustering such that each pair of clusters is linearly separable.*
- *The problem is solvable by an algorithm in polynomial time.*

Thank you very much!