Size constrained clustering problems in fixed dimension

Department of Computer Science
University of Milan

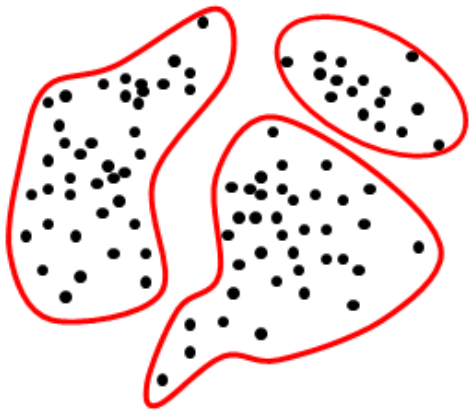
Author
Jianyi LIN

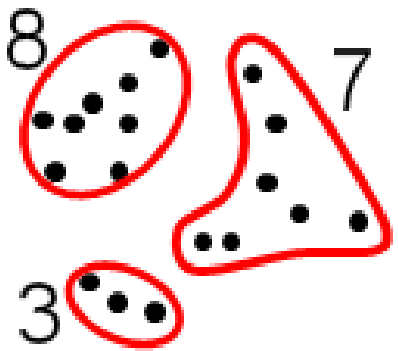
ICTCS 2012 - Varese 19th Sep 2012

Outline

Problem and motivation

- Clustering or cluster analysis: classical method in statistics and unsupervised machine learning.
- Clustering: grouping objects into "significant classes" with respect to a similarity measure
- k-Means or sum-of-squares clustering: NP-hard in arbitrary dimension [Aloise et al. 2009] or with arbitrary k [Mahajan et al. 2009, Vattani 2009].





- Incorporating a priori information on the clusters into the algorithms can increase clustering performance [Wagstaff & Cardie 2000, Bradley et al. 2000, Tung et al. 2001]
- A priori information $\Longrightarrow$ Constrained clustering
- In this work: constraints on the cardinalities [Zhu et al. 2010]

Definitions

Fixed $\| \ \|_p$ $(p \geq 1)$:

$X = \{x_1, ..., x_n\} \subset \mathbb{R}^d$

- cluster $:= A \subseteq X$
- p-centroid $:= C_A = \text{argmin}_{\mu \in \mathbb{R}^d} \sum_{x \in A} \|x - \mu\|_p^p$
- cost $W(A_i) := \sum_{x \in A_i} \|x - C_{A_i}\|_p^p$

- k-clustering $:= k$-partition $\{A_1, ..., A_k\}$ of X
- cost of the clustering $:= W(A_1, ..., A_k) = \sum_i W(A_i)$

Preliminary definitions

Size Constrained Clustering (SCC)

Instance:
- positive integer d
- positive integer k
- $X = \{x_1, ..., x_n\} \subset \mathbb{R}^d$
- positive integers $m_1, ..., m_k$ s.t. $\sum m_i = n$

Admissible solutions: k-partition $\{A_1, ..., A_k\}$ of X with $|A_j| = m_j$

Cost: $W(A_1, ..., A_k)$

Type: min

Fixing k or d: k-SCC, SCC-d, k-SCC-d

Preliminary definitions

Relaxed Constraints Clustering (RCC)

Instance:
- positive integer d
- positive integer k
- $X = \{x_1, ..., x_n\} \subset \mathbb{R}^d$
- set positive integers $\mathcal{M} = \{m_1, ..., m_\ell\}$

Admissible solutions: k-partition $\{A_1, ..., A_k\}$ of X with $|A_j| \in \mathcal{M}$

Cost: $W(A_1, ..., A_k)$

Type: min

REMARK:

- RCC with $\mathcal{M} = \{1, ..., n\}$ corresponds to the classical (unconstrained) clustering.
- When $k = 2$, the SCC problem and the RCC problem with $|\mathcal{M}| = 2$ are equivalent.

Results (1)

Previous results [Bertoni et al. 2012]:

- k-SCC with $k = 2$ and $m = \frac{n}{2}$ is NP-hard for all norm $\| \ \|_p$
- SCC-d is NP-hard even if $d = 1$

Here we present:

- RCC-d is NP-hard even if $d = 2$. Polynomial in case $d = 1$
- Evidence that for non-integer rational p the problem cannot be solved in polynomial time

Results (2)

The problem could be solved in polynomial time only when:

- d, k are fixed
- p is integer

- For $p = 2$, SCC can be solved in the plane with $k = 2$ in time $O(n\sqrt[3]{m} \cdot \log^2 n)$.
- For $p = 2$, all the SCC problems, with $m = 1, ..., \lfloor n/2 \rfloor$, can be solved at once by an algorithm working in time $O(n^2 \cdot \log n)$.
- For integer $p > 2$, RCC is solved in fixed dimension with $k = 2$ in polynomial time (even with p coded in unary in the instance).

RCC-d is NP-hard

Consider the norm $\| \ \|_2$.

Theorem

RCC-d is NP-hard even if $d = 2$.

This result is obtained as a consequence of the reduction:

$$\text{Planar 3-SAT} <_P \{2,3\}\text{-RCC}$$

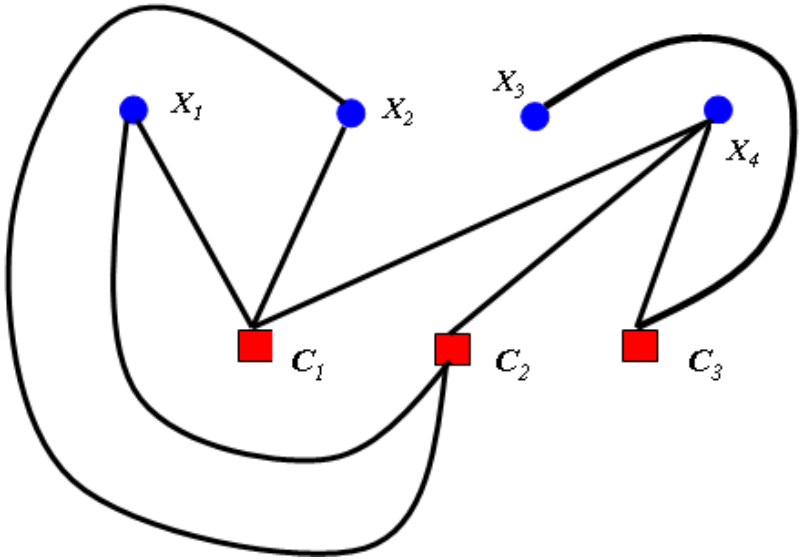
Problem: $\{2,3\}$-**RCC**

Instance:
- $X = \{x_1, ..., x_n\} \subset \mathbb{Q}^2$
- positive integer k
- positive rational λ

Question: Is there a partition $\{A_1, ..., A_k\}$ of X s.t. $W(A_1, ..., A_k) \leq k$ and $|A_i| \in \{2,3\}$?

Planar 3-SAT problem

A Planar 3-CNF $\Phi = \underbrace{(x_1 \vee \bar{x}_2 \vee x_4)}_{C_1} \wedge \underbrace{(\bar{x}_1 \vee x_2 \vee x_4)}_{C_2} \wedge \underbrace{(\bar{x}_2 \vee x_3 \vee \bar{x}_4)}_{C_3}$



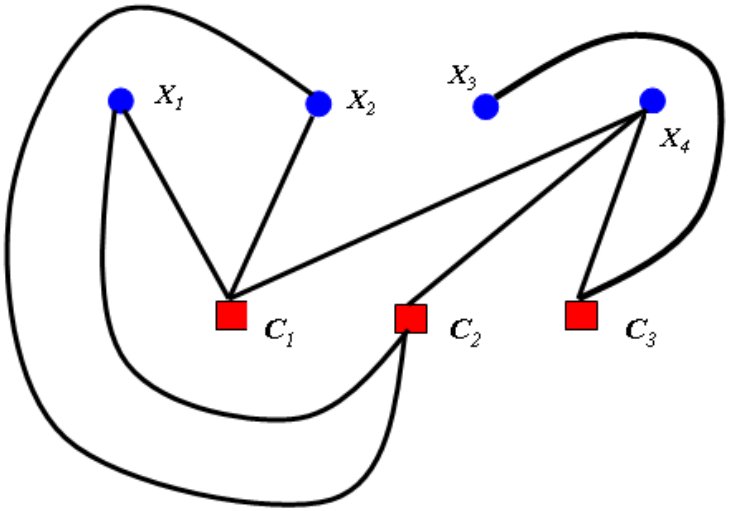
Problem: Planar 3-SAT
Instance: Planar 3-CNF Φ
Question: Is Φ satisfiable?

In the example above Φ is satisfied by $(x_1, x_2, x_3, x_4) = (F, F, T, T)$.

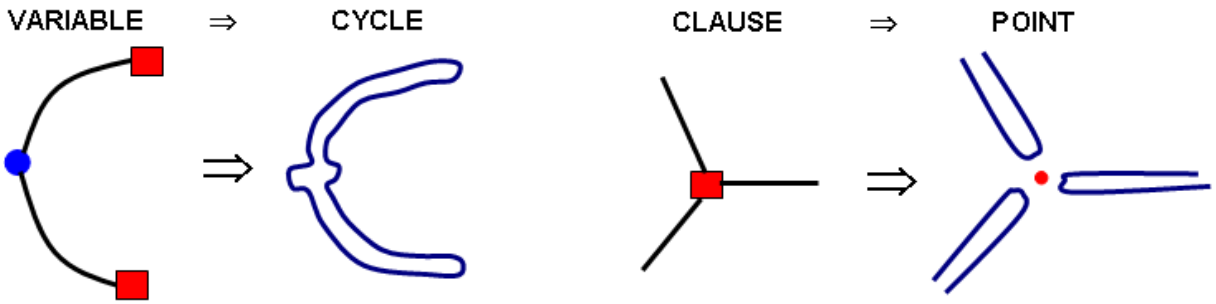
Proof of the reduction

$$\Phi = \underbrace{(x_1 \vee \bar{x}_2 \vee x_4)}_{C_1} \wedge \underbrace{(\bar{x}_1 \vee x_2 \vee x_4)}_{C_2} \wedge \underbrace{(\bar{x}_2 \vee x_3 \vee \bar{x}_4)}_{C_3}$$

Variables x_1, x_2, x_3, x_4, clauses $C_1, C_2, C_3 \Longrightarrow$ planar graph G:



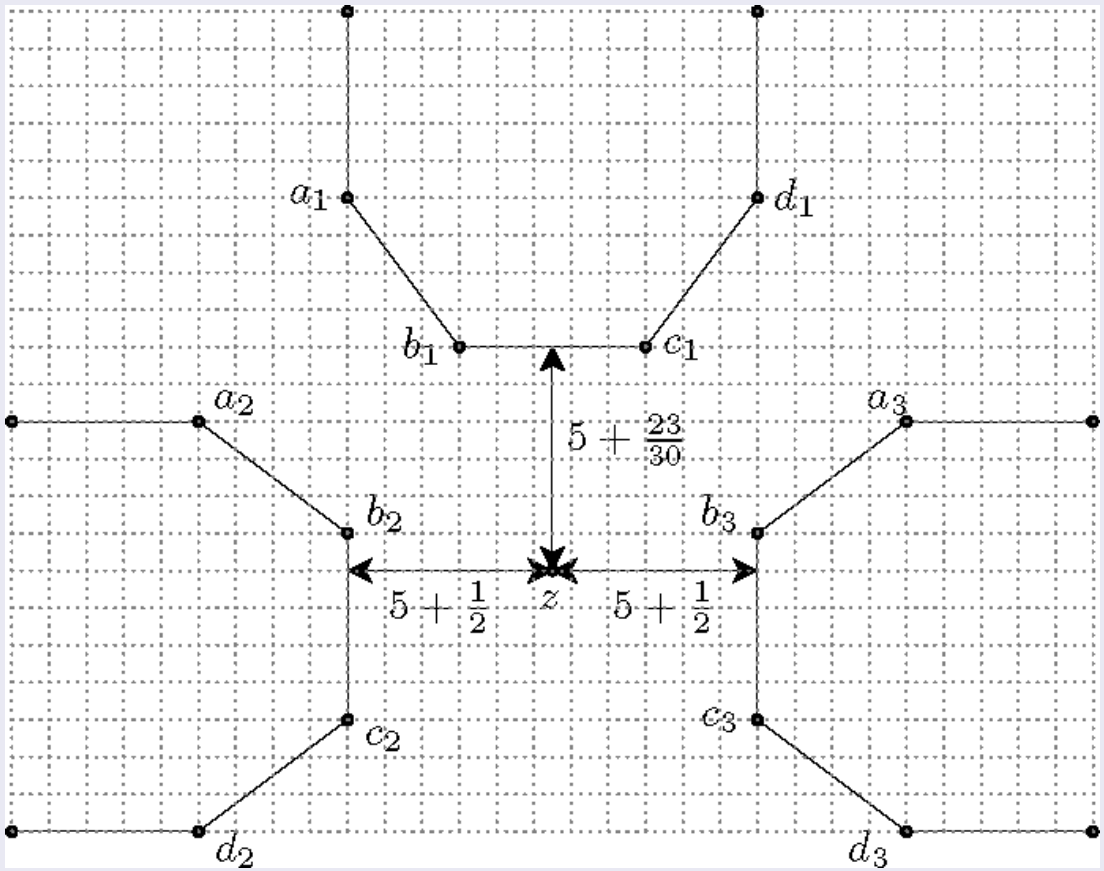
Embedded in rational coordinate grid



Embedding some points into rational coordinate grid

Lemma

Consider the points in figure below and let $A = \{z, b_1, c_1\}$, $B = \{z, b_2, c_2\}$. *We have that* $W(A) = W(B) = 23402/675$.



Localisation of p-centroid

When p is a non-integer rational we can see that also the minor problem of localising the centroid of a set of integers is far from being easy.

- p-LC Problem: given integers $x_1, ..., x_n$ and an integer h, decide whether the p-centroid of $\{x_1, ..., x_n\}$ is $> h$.
- SQRT-Sum Problem: requires to decide, given positive integers $a_1, ..., a_q, b_1, ..., b_r$, whether $\sqrt{a_1} + ... + \sqrt{a_r} > \sqrt{b_1} + ... + \sqrt{b_r}$.

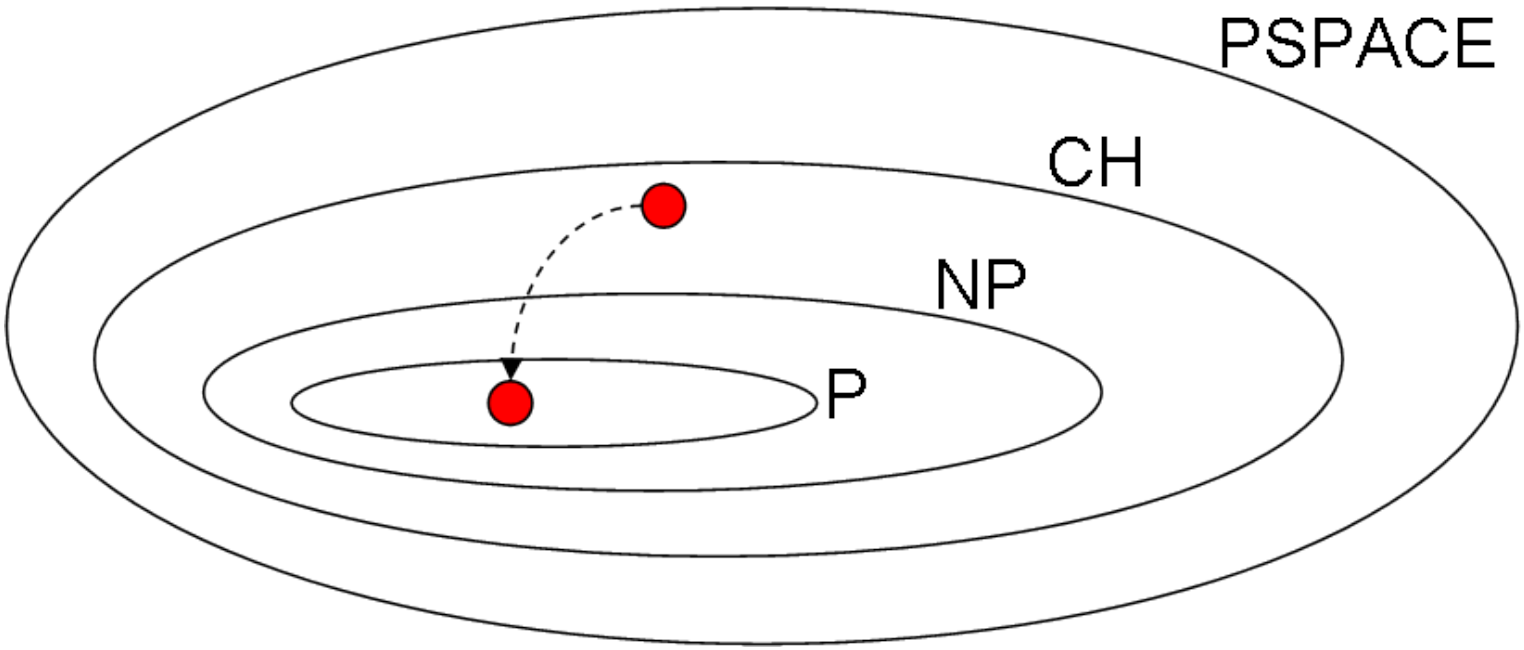
Theorem

SQRT-Sum is polynomially reducible to $\frac{3}{2}$-LC.

SQRT-Sum

OPEN PROBLEM:
Is SQRT-Sum in NP? [Garey et al. 1976]

Best result: SQRT-Sum $\in$ CH [Allender et al. 2006]



Subproblems

Problem

GENERAL CLUSTERING: Hard

What happens when

- *d is fixed* *(e.g. $d = 2$)*
- *k is fixed* *(e.g. $k = 2$)*

?

Fixing $k = 2$ is particularly interesting for being the major step repeated in the divisive hierarchical clustering methods.

Size constrained 2-clustering on the plane

Case:

$\mathbb{R}^2$ endowed with $\| \ \|_2$

Theorem

The SCC problem in the plane with $k = 2$ and constraint $|A| = m$ is solvable in $O(n \cdot \sqrt[3]{m} \cdot \log^2 n)$ time.

REMARK: The problem is equivalent to RCC in the plane with $k = 2$, $\mathcal{M} = \{m, n - m\}$.

The proof is based on:

1. Separation Property
2. Efficient dynamic data structures for Convex Hull:
   - $O(n \log^2 n)$ [Overmars & Van Leeuwen 1981]
   - $O(n \log n)$ amortized time [Chan 2001]
3. Upper bound $O(n\sqrt[3]{k})$ for the number of k-sets of n points on the plane [Erdös et al. 1973, Dey 1998]

Separation Property

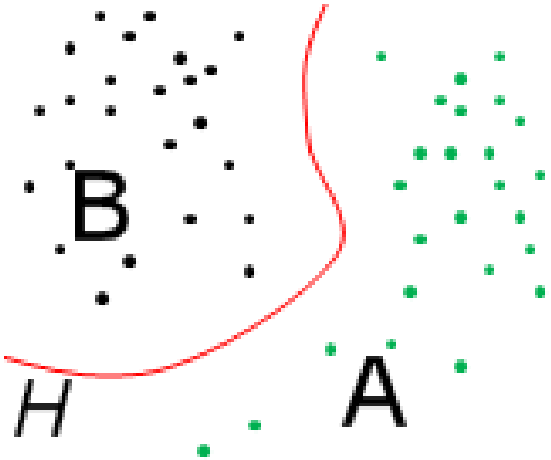
Theorem

Consider the norm $\| \ \|_p$ *with integer* $p > 1$. *Let*

- $X = \{x_1, ..., x_n\} \subset \mathbb{R}^d$, *and*
- $\{A, B\}$ *be the optimal solution of SCC with point set* X, $k = 2$ *and constraint* $|A| = m$

Then there exists $c \in \mathbb{R}$ *such that:*

$$\forall x_i \in A \qquad \|x_i - C_A\|_p^p - \|x_i - C_B\|_p^p < c$$
$$\forall x_j \in B \qquad \|x_j - C_A\|_p^p - \|x_j - C_B\|_p^p > c$$



That is, A and B are separated by a hypersurface

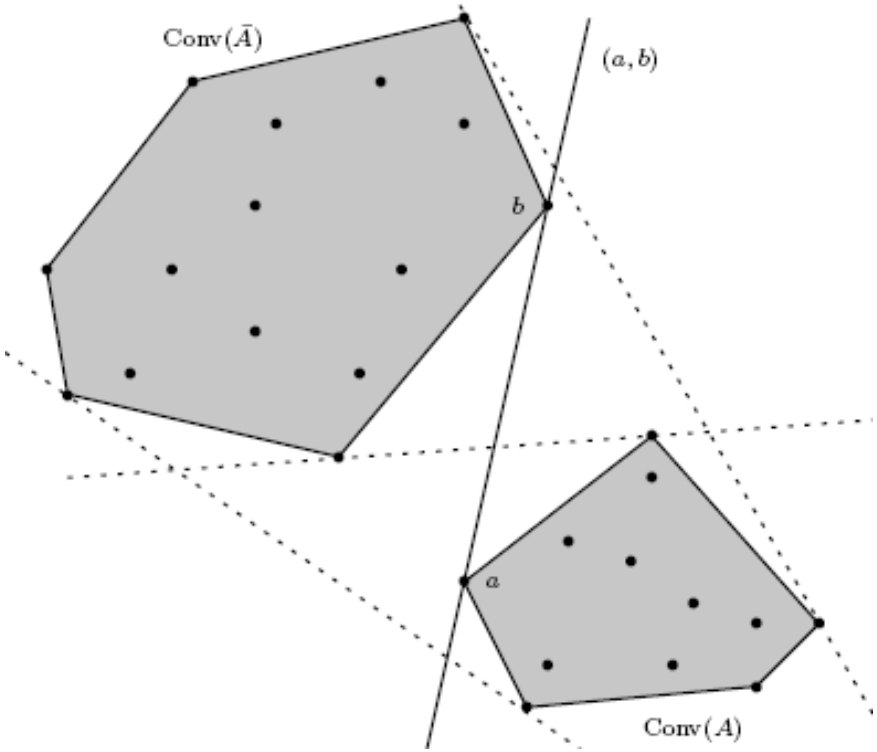
$$H : \|x - a\|_p^p - \|x - b\|_p^p - c = 0$$

Particular cases:

- $d = 1 \Longrightarrow$ String Property [Fisher 1958].
- $d = 2, p = 2 \Longrightarrow$ Separation with straight lines!

Constructing the m-sets in efficient manner

Idea of the algorithm:



- Convex hulls Conv(A) and Conv($\bar{A}$), and a bitangent (a, b).
- Find the next bitangent and swap two points $c \in A$, $d \in \bar{A}$, thus obtaining new clusters A' and $\bar{A}'$
- Calculate the cost of the new 2-clustering $W(A', \bar{A}')$ with direct formula
- Reiterating those abstract operations corresponds to visiting all the possible m-sets of X

Bound on the abstract operations and dynamic data structures

DEF.: Given $X = \{x_1, ..., x_n\} \subset \mathbb{R}^2$, a *m-set* is a subset $A \subset X$ with $|A| = m$ s.t. A and $\bar{A}$ are separable by a straight line.

OPEN QUESTION: How many? [Erdös 1973]

- Upper bound:
  $O(n\sqrt{m})$ [Erdös 1973]
  $O(n\sqrt[3]{m})$ [Dey 1998]
- Lower bound:
  $ne^{\Omega(\sqrt{\log m})}$ [Tóth 2001]

Operations on efficient dynamic data structures for convex hulls:

- Only INSERTION: $O(\log n)$ [Preparata 1979]
- INSERTION+DELETION: $O(\log^2 n)$ [Overmars & Van Leeuwen 1981]
- INSERTION+DELETION: $O(\log n)$ amortized time [Chan 2001]

Full planar 2-RCC

Instead of solving RCC in the plane with $k = 2$ for a particular value of the constraint $|A| = m$, we may ask to solve the RCC problem in the plane with $k = 2$ for all possible constraints $|A| = 1, ..., \lfloor n/2 \rfloor$. Such a problem can be called Full 2-RCC.

Theorem

There is an algorithm for solving Full 2-RCC in the plane in time $O(n^2 \cdot \log n)$.

- This algorithm outputs all the optimal 2-clusterings $\{A_m, \bar{A}_m\}$ of X with constraint $|A_m| = m$, for all m, $1 \leq m \leq \lfloor n/2 \rfloor$.
- It is based on a smart enumeration of the pairs of points in X, which correspond to the straight lines separating the clusters, as stated in the Separation Property:
  - for every point $x \in X$: order $X \setminus \{x\}$ w.r.t. angle (with pole in x), then enumerate all separating straight lines (x, y), with $y \in X \setminus \{x\}$ w.r.t. this order;
  - cost of 2-clustering computed in $O(1)$ time.
  - $\therefore O(n^2 \log n)$

What if $p > 2$

Case: p is even

Separating curve:

$(x - a_1)^p + (y - a_2)^p - (x - b_1)^p - (y - b_2)^p - c = 0$

- Yields a polynomial in variables x, y and parameters a_1, a_2, b_1, b_2, c.
- The description can be generalized for $d > 2$.

Formulation of the size constrained clustering problem in Real Algebraic Geometry.

Main tool: CAD [Collins 1975]. CAD's applications: quantifier elimination in 1^{st}-order theory of reals, robot inverse kinematic.

2-clustering problem with even p

Theorem

The 2-SCC problem with size constraints m, $1 \leq m \leq \lfloor n/2 \rfloor$, in fixed dimension d with norm $\| \ \|_p$, even integer p, can be solved in polynomial time w.r.t. to the input size and p

Proof Sketch: $X \ni x_i \longmapsto p_i \in \mathbb{R}[\alpha]$ in the surface's parameters $\alpha = (\mu_1, ..., \mu_d, \lambda_1, ..., \lambda_d, \gamma)$:

$$p_i(\mu, \lambda, \gamma) = \|x_i - \mu\|_p^p - \|x_i - \lambda\|_p^p - \gamma \qquad F = \{p_i \in \mathbb{R}[\alpha] : x_i \in X\}$$

$$F \Longrightarrow \mathbb{R}^{2d+1} = \underbrace{R_1 \sqcup ... \sqcup R_t} \qquad \text{with} \qquad t = (np)^{O(1)^d}$$

$$R_j\text{'s are semi-algebraic} \qquad \operatorname{sgn}(p_i(R_j)) \in \{+1, -1\}$$

- CAD algorithm constructs: representatives $\bar{\alpha}_j \in R_j \cap \mathbb{Q}^{2d+1}$ using
  - theory of elimination through resultant
  - Sturm sequences and root bounds to isolate algebraic roots

  $$\bar{\alpha}_j \longmapsto \text{clustering } \{A_j, \bar{A}_j\} \text{ of } X$$
- Comparison of two 2-clusterings by a numerical approximation technique exploiting Canny's Gap.

2-clustering problem with odd p

Case: p is odd

Problem: the separating hypersurface is no longer algebraic. Nonetheless, the case of odd p can be directly reduced to the case of even p by easily enriching the collection $F = \{p_1, ..., p_n\}$:

$$p_i(\mu, \lambda, \gamma) = \|x_i - \mu\|_p^p - \|x_i - \lambda\|_p^p - \gamma$$

- By eliminating absolute values:

$$p_i \longmapsto \Psi_i = \{\bar{p}_{i\sigma_i\tau_i} \in \mathbb{R}[\alpha] : \sigma_i, \tau_i \in \{+1, -1\}^d\}$$

- $G := \{\text{polynomials } (x_{i\ell} - \mu_\ell), (x_{i\ell} - \lambda_\ell)\}$ arguments of absolute values

$$H = G \cup \Psi_1 \cup ... \cup \Psi_n$$

This allows one to take into account the further decomposition of the parameter space $\mathbb{R}^{2d+1}$ due to the absolute value function.
Prop.: A decomposition of $\mathbb{R}^{2d+1}$ adapted to H is also adapted to F.

Theorem

The 2-SCC problem with size constraints m, $1 \leq m \leq \lfloor n/2 \rfloor$, in fixed dimension d with norm $\| \|_p$, odd integer p, can be solved in polynomial time w.r.t. the input size and p.

Final words

- Size constrained clustering is difficult in general.
- Instead, fixing # of clusters and dimension yields polynomial-time algorithms.
- SCC with euclidean norm can be tackled by methods of combinatorial geometry.
- When $p > 2$ the SCC can be studied within the frame of real algebraic geometry, but not complex algebraic geometry.
- Open problems: avoid CAD; approximation technique for constrained clustering; heuristic for constrained clustering;
- Recent considerations for avoiding CAD

Questions?

Thank you for your attention!

Canny's Gap

Theorem (Canny's Gap)

Let $(x_1, ..., x_N)$ be a solution of an algebraic system of N equations in N unknowns having a finite number of solutions, with maximum degree d and with coefficients in $\mathbb{Z}$ smaller or equal to M in absolute value. Then for each $i = 1, ..., N$:

$$\text{either } x_i = 0 \text{ or } |x_i| > (3Md)^{-Nd^N}$$

- Powerful tool for numerically solving symbolic decision problems
- Decide whether an algebraic number represented by algebraic equation is null