$k-$means clustering

through continuous optimization

Jacob Kogan
Department of Mathematics & Statistics
and
Department of CSEE
University of Maryland, Baltimore County (UMBC)
Baltimore, MD 21250, USA

Partitions

$\mathcal{A} = \{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$ is a set of vectors in $\mathbf{R}^n$.

A partition Π of $\mathcal{A}$ is

$$\Pi = \{\pi_1, \ldots, \pi_k\}$$

$$\pi_1 \cup \ldots \cup \pi_k = \{\mathbf{a}_1, \ldots, \mathbf{a}_m\}, \text{ and } \pi_i \bigcap \pi_j = \emptyset \text{ if } i \neq j.$$

q is a real valued function whose domain is the set of subsets of $\mathcal{A}$.

The quality of the partition is given by

$$Q(\Pi) = q(\pi_1) + \ldots + q(\pi_k).$$

What do we want?

To identify an optimal partition

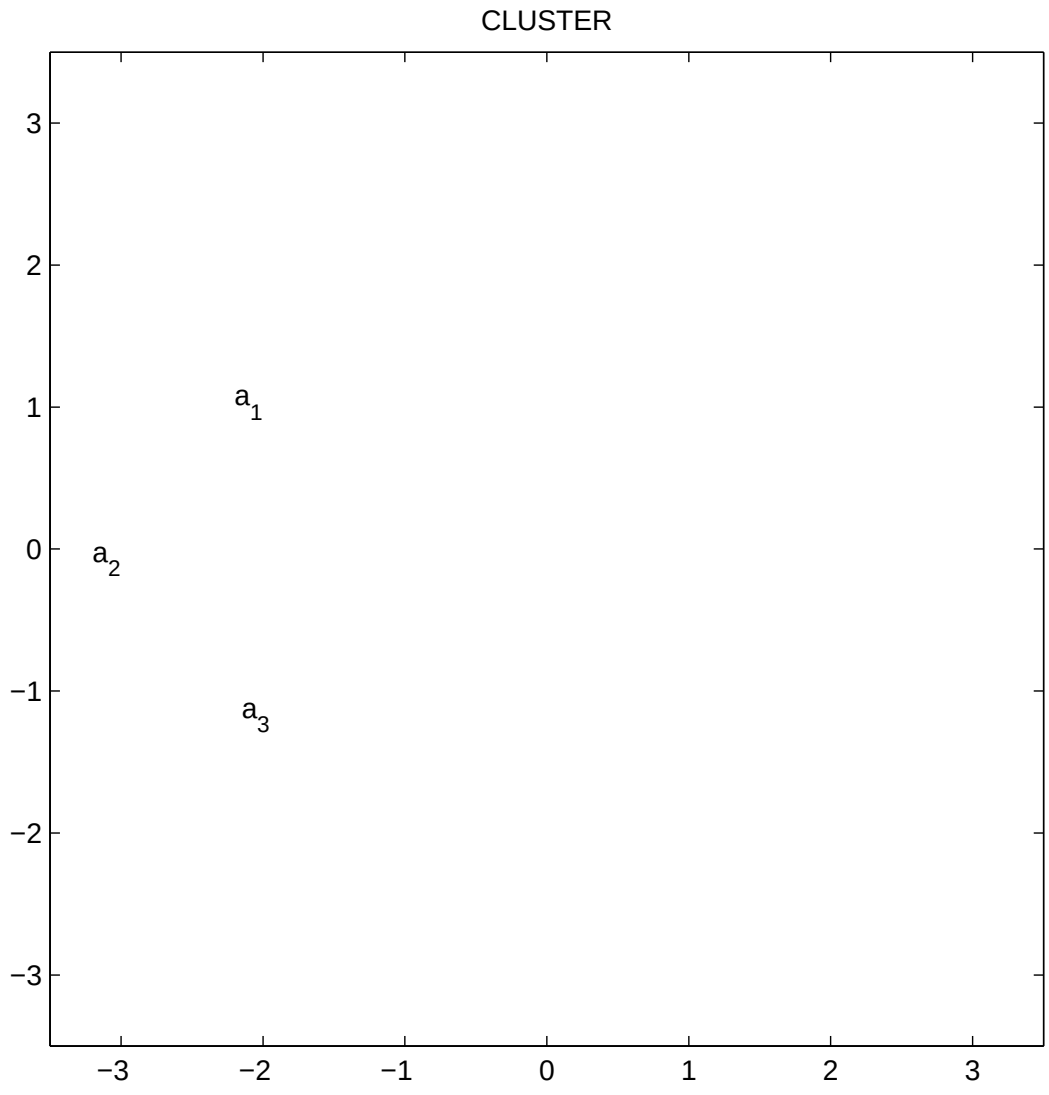
$$\Pi^o = \{\pi_1^o, \ldots, \pi_k^o\},$$

i.e., one that optimizes

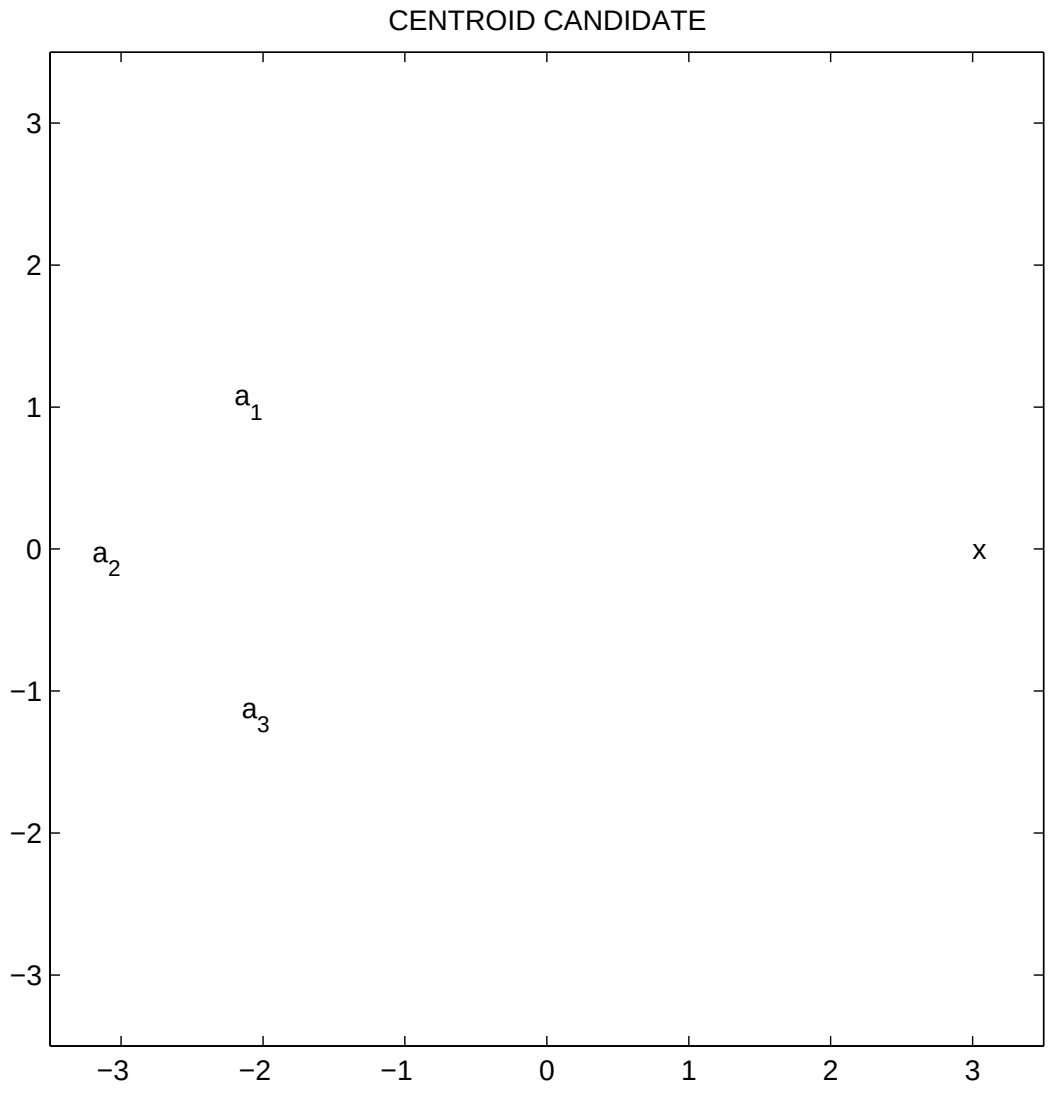
$$Q(\Pi) = q(\pi_1) + \ldots + q(\pi_k).$$

In general the solution is available when the dimension of the vector space is **ONE**.

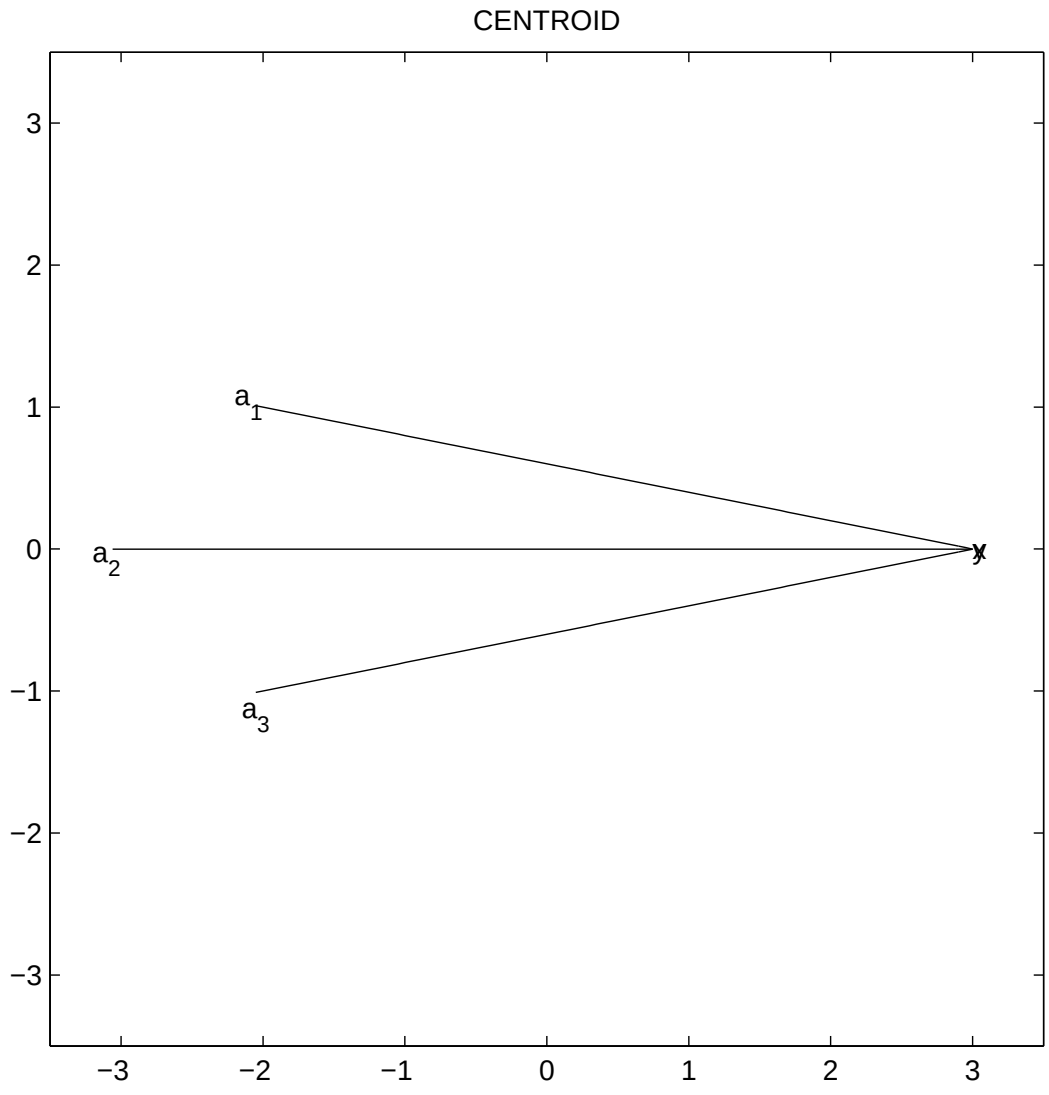
Centroid



Centroid



Centroid



$$f(\mathbf{x}) = d(\mathbf{x}, \mathbf{a}_1) + d(\mathbf{x}, \mathbf{a}_2) + d(\mathbf{x}, \mathbf{a}_3)$$

centroid-cluster quality association

$$\mathbf{c} = \mathbf{c}(\pi) = \arg\min \left\{ \sum_{\mathbf{a} \in \pi} d(\mathbf{x}, \mathbf{a}),\ \mathbf{x} \in \mathbf{C} \right\}.$$

If the quality $q(\pi)$ of a cluster π is defined by

$$q(\pi) = \sum_{\mathbf{a} \in \pi} d(\mathbf{c}(\pi), \mathbf{a}),$$

then centroids and partitions can be associated.

centroid-partition association

1. For a set of centroids $\{\mathbf{c}_1, \ldots, \mathbf{c}_k\}$ define a partition $\{\pi_1, \ldots, \pi_k\}$ of the set $\mathcal{A}$ by:

$$\pi_i = \{\mathbf{a} \mid d(\mathbf{c}_i, \mathbf{a}) \leq d(\mathbf{c}_j, \mathbf{a}) \text{ for each } j \neq i\}$$

   (we break ties arbitrarily).

2. Given a partition $\{\pi_1, \ldots, \pi_k\}$ of the set $\mathcal{A}$ define the corresponding centroids $\{\mathbf{c}_1, \ldots, \mathbf{c}_k\}$ by

$$\mathbf{c}_i = \arg\min \left\{ \sum_{\mathbf{a} \in \pi_i} d(\mathbf{x}, \mathbf{a}),\ \mathbf{x} \in \mathbf{C} \right\}.$$

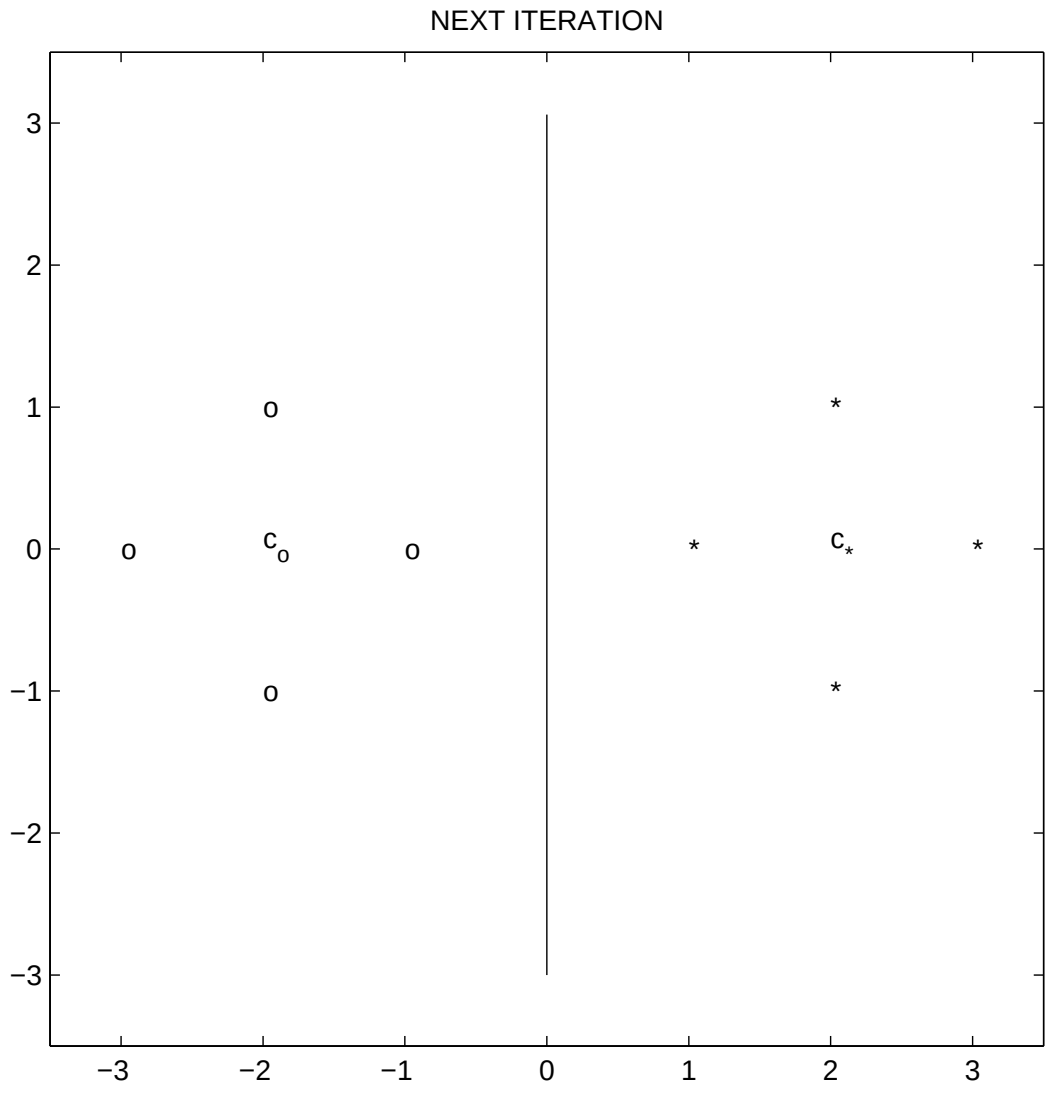
Squared Euclidean Distance

We shall focus now on the “distance like" function

$$d(\mathbf{x}, \mathbf{a}) = \|\mathbf{x} - \mathbf{a}|^2.$$

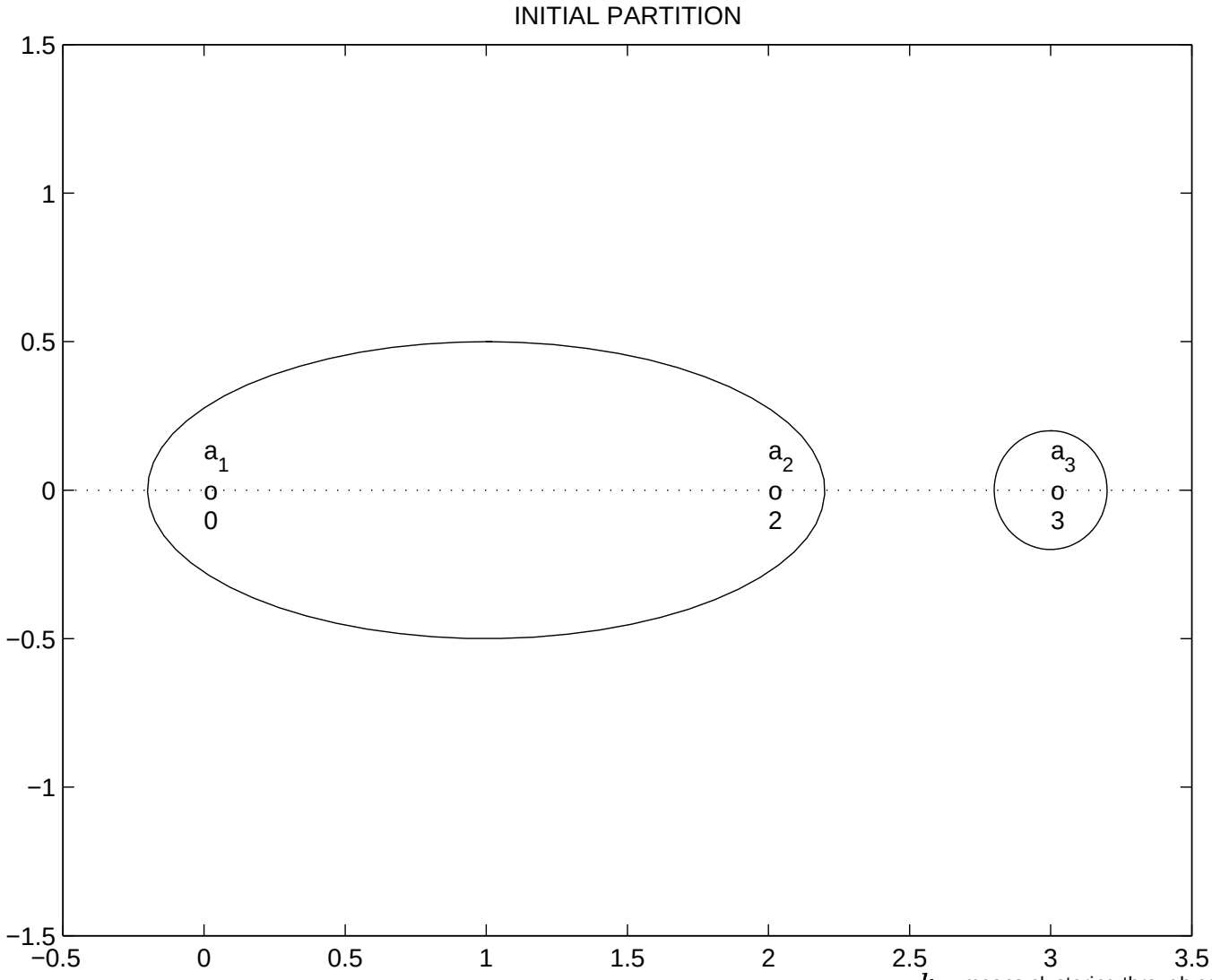
$$\arg\min \left\{ \sum_{i=1}^{m} \|\mathbf{x} - \mathbf{a}_i|^2 \right\} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{a}_i.$$

convexity of final partition



"trapped" k–means

$$\mathcal{A} = \{0, 2, 3\},\ \pi_1^{(0)} = \{0, 2\}\,,\ \pi_2^{(0)} = \{3\}.$$



Enhanced k-means algorithm

1. Set $t = 0$.
2. Start with an arbitrary partitioning $\Pi^{(t)} = \left\{\pi_1^{(t)}, \ldots, \pi_k^{(t)}\right\}$.
3. Run batch k-means until no vector movement is detected.
4. Run one iteration of incremental k-means.
   if (vector movement is detected) go to Step 3.
5. Stop.

Cost of an Incrimental Iteration

The exact change in the value of the objective function caused by removing $\mathbf{a}$ from π_i and assigning it to π_j is

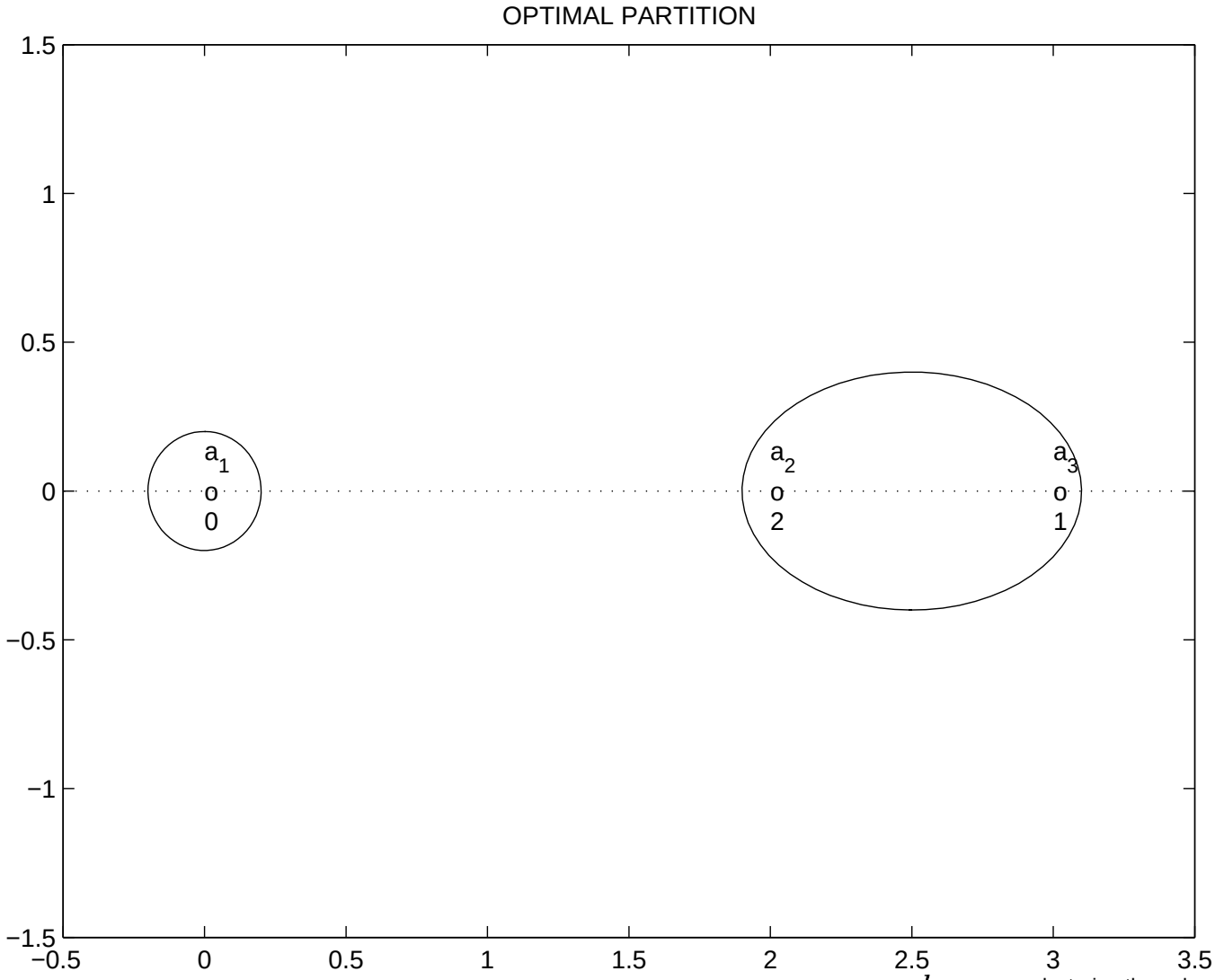
$$\Delta_{\mathrm{I}} = -\frac{m_i}{m_i-1}\left\|\mathbf{a}-\mathbf{c}(\pi_i)\right\|^2 + \frac{m_j}{m_j+1}\left\|\mathbf{a}-\mathbf{c}(\pi_j)\right\|^2$$

$$\Delta_{\mathrm{B}} = -\left\|\mathbf{a}-\mathbf{c}(\pi_i)\right\|^2 + \left\|\mathbf{a}-\mathbf{c}(\pi_j)\right\|^2$$

Batch $k-$means moves $\mathbf{a}$ from π_i to π_j when $\Delta_{\mathrm{B}} < 0$.

optimal partition by enhanced k–means

$$\mathcal{A} = \{0, 2, 3\},\ \pi_1^{(1)} = \{0\}\,,\ \pi_2^{(1)} = \{2, 3\}.$$



Continuous Optimization Problem

$$Q(\Pi) = q(\pi_1) + \ldots + q(\pi_k).$$

$$\mathbf{x}_i = \mathbf{c}(\pi_i)$$

To identify an optimal partition

$$\Pi^o = \{\pi_1^o, \ldots, \pi_k^o\},$$

one has to find

$$\mathbf{x}^o = (\mathbf{x}_1^o, \ldots, \mathbf{x}_k^o)^T \in \mathbf{R}^{nk}$$

that solves

$$\min_{\mathbf{x} \in \mathbf{R}^N} F(\mathbf{x}) = \sum_{i=1}^{m} \min_{1 \le l \le k} \|\mathbf{x}_l - \mathbf{a}_i\|^2.$$

Continuous Optimization Problem

- Rose, Gurewitz and Fox–“A deterministic annealing approach to clustering", 1990.
- Nasraoui and Krishnapuram–“Crisp Interpretations of Fuzzy and Possibilistic Clustering Algorithms", 1995.
- Zhang, Hsu and Dayal–“K-Harmonic Means–A Data Clustering Algorithm", 1999.

Continuous Optimization Problem

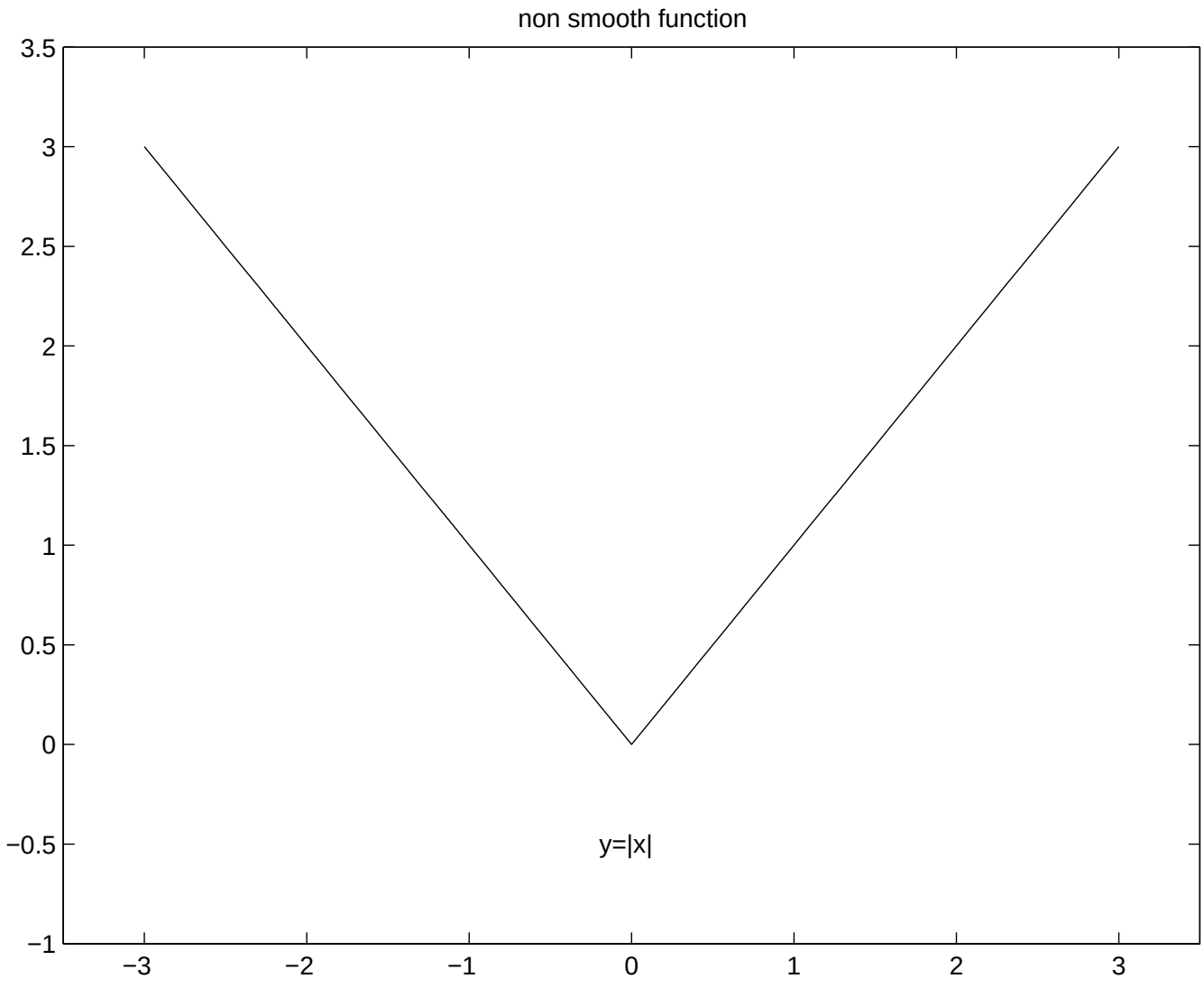
The problem

$$\min_{\mathbf{x}\in\mathbf{R}^N} F(\mathbf{x}) = \sum_{i=1}^{m} \min_{1\leq l\leq k} \|\mathbf{x}_l - \mathbf{a}_i\|^2$$

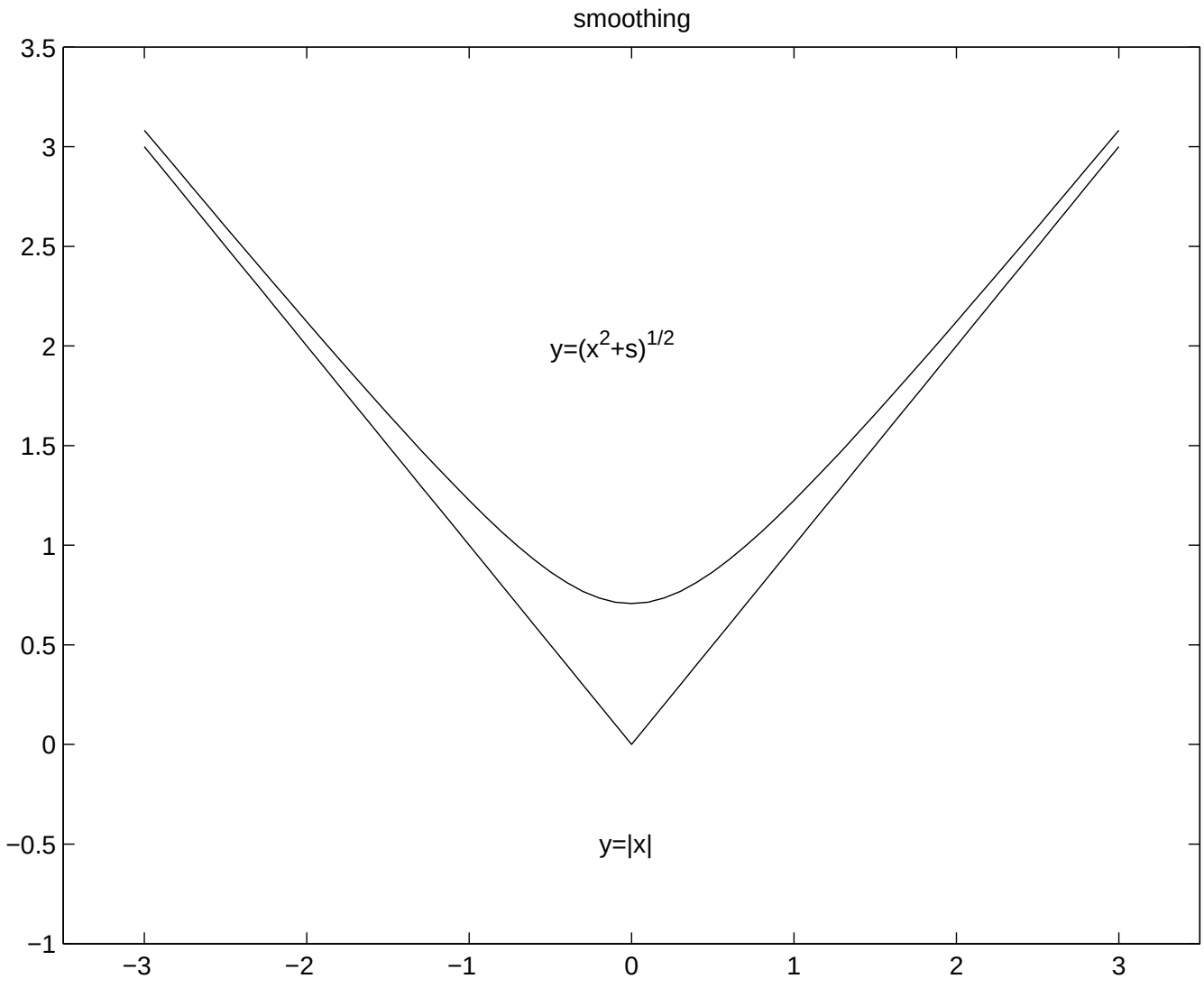
is:

1. non convex
2. non smooth

Smoothing



Smoothing



Smoothing

For any two numbers a and b one has

$$\max\{a, b\} = \lim_{s \to 0^+} s \log \left(e^{\frac{a}{s}} + e^{\frac{b}{s}} \right)$$

and

$$\min\{a, b\} = \lim_{s \to 0^+} -s \log \left(e^{-\frac{a}{s}} + e^{-\frac{b}{s}} \right).$$

Smooth Approximation

Substitute

$$\min_{\mathbf{x}\in\mathbf{R}^N} F(\mathbf{x}) = \sum_{i=1}^{m} \min_{1\le l\le k} \|\mathbf{x}_l - \mathbf{a}_i\|^2$$

by

$$\min_{\mathbf{x}\in\mathbf{R}^N} F_s(\mathbf{x}) = \sum_{i=1}^{m} -s \log\left(\sum_{l=1}^{k} e^{-\frac{\|\mathbf{x}_l-\mathbf{a}_i\|^2}{s}}\right)$$

Smooth Approximation

so that

$$\lim_{s \to 0^+} F_s(\mathbf{x}) = F(\mathbf{x})$$

and F_s approximates F uniformly

$$0 \leq F(\mathbf{x}) - F_s(\mathbf{x}) \leq sm \log k.$$

Optimality Condition

Fix $s > 0$ and solve

$$\min_{\mathbf{x} \in \mathbf{R}^N} F_s(\mathbf{x}) = \sum_{i=1}^{m} -s \log \left(\sum_{l=1}^{k} e^{-\frac{\|\mathbf{x}_l - \mathbf{a}_i\|^2}{s}} \right).$$

The necessary local optimality condition is

$$\nabla F_s(\mathbf{x}) = 0.$$

Optimality Condition

$$\sum_{i=1}^{m}(\mathbf{x}_l - \mathbf{a}_i)\frac{e^{-\frac{\|\mathbf{x}_l-\mathbf{a}_i\|^2}{s}}}{\sum_{j=1}^{k} e^{-\frac{\|\mathbf{x}_j-\mathbf{a}_i\|^2}{s}}} = 0 \qquad \text{for each } l = 1, \ldots, k.$$

Denote

$$\frac{e^{-\frac{\|\mathbf{x}_l-\mathbf{a}_i\|^2}{s}}}{\sum_{j=1}^{k} e^{-\frac{\|\mathbf{x}_j-\mathbf{a}_i\|^2}{s}}} \text{ by } \rho^{il}(\mathbf{x}, s)$$

and get

$$\sum_{i=1}^{m}(\mathbf{x}_l - \mathbf{a}_i)\rho^{il}(\mathbf{x}, s) = 0.$$

Optimality Condition

$$\mathbf{x}_l = \sum_{i=1}^{m} \frac{\rho^{il}(\mathbf{x}, s)}{\sum_{j=1}^{m} \rho^{jl}(\mathbf{x}, s)} \mathbf{a}_i = \sum_{i=1}^{m} \lambda^{il}(\mathbf{x}, s) \mathbf{a}_i$$

with

$$\lambda^{il}(\mathbf{x}, s) > 0, \text{ and } \sum_{i=1}^{m} \lambda^{il}(\mathbf{x}, s) = 1.$$

smoka

Pick $\texttt{tol} > 0$, smoothing parameter $s > 0$, and k initial centroids $\mathbf{x}(0) = (\mathbf{x}_1^T(0), \ldots, \mathbf{x}_k^T(0))^T \in \mathbf{R}^N$.
Do the following:

1. Set $t = 0$.
2. For each $l = 1, \ldots, k$ compute

$$\mathbf{x}_l(t+1) = \sum_{i=1}^{m} \lambda^{il}(\mathbf{x}(t), s)\mathbf{a}_i.$$

3. If $F_s(\mathbf{x}(t)) - F_s(\mathbf{x}(t+1)) > \texttt{tol}$

   set $t = t + 1$

   goto step 2
4. Stop.

DATA

classic3–a merger of the three document collections available from `http://www.cs.utk.edu/~lsi/`:

- DC0 (Medlars Collection, 1033 medical abstracts).
- DC1 (CISI Collection, 1460 information science abstracts).
- DC2 (Cranfield Collection, 1398 aerodynamics abstracts).

Initial Partition

	DC0	DC1	DC2
cluster 0	907	91	13
cluster 1	120	7	1372
cluster 2	6	1362	13
“empty" documents			
cluster 3	0	0	0

Table 1: Collection: **classic3**. PDDP generated initial “confusion" matrix with **250** “misclassifi ed" documents using **600** best terms, $Q = 3612.61$

performance

algorithm	iterations	misclass	Q
PDDP		250	3612.6
batch k–means	3	131	3608.1
k–means	87	79	3605.5
smoka	7	73	3605.5

Table 2: Collection: **classic3**. Number of iterations, misclassifi cations, and partition quality per clustering algorithm applied to the initial partition generated by PDDP, the vector space dimension is **600**

DATA

The 20 newsgroups dataset available at `http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html`.

- The “mini" dataset is a subset of the “full" dataset with 100 documents from each of the 20 Usenet newsgroups.
- The “full" dataset of 19997 messages taken from 20 Usenet newsgroups.

performance

algorithm	iterations	Q
PDDP		1758.4
batch k–means	11	1737.7
k–means	473	1721.9
smoka	15	1726.3

Table 3: Collection: **the "mini" subset of the 20 newsgroups dataset.** Number of iterations per clustering algorithm applied to the initial partition generated by PDDP, 2000 vectors of dimension **600**

performance

algorithm	iterations	Q
PDDP		18156
batch k–means	47	17956
k–means	5862	17808
`smoka`	51	17810

Table 4: Collection: **the "full" 20 newsgroups dataset.** Number of iterations per clustering algorithm applied to the initial partition generated by PDDP, 19997 vectors of dimension 1000 from the "full" 20 newsgroups dataset