Lecture 11

Clustering, Robustness and Sparsity

Bao-Jun Cai, 5/13/2026

Introduction to Algorithms for Data Science and Physics IMP@Fudan, 2026

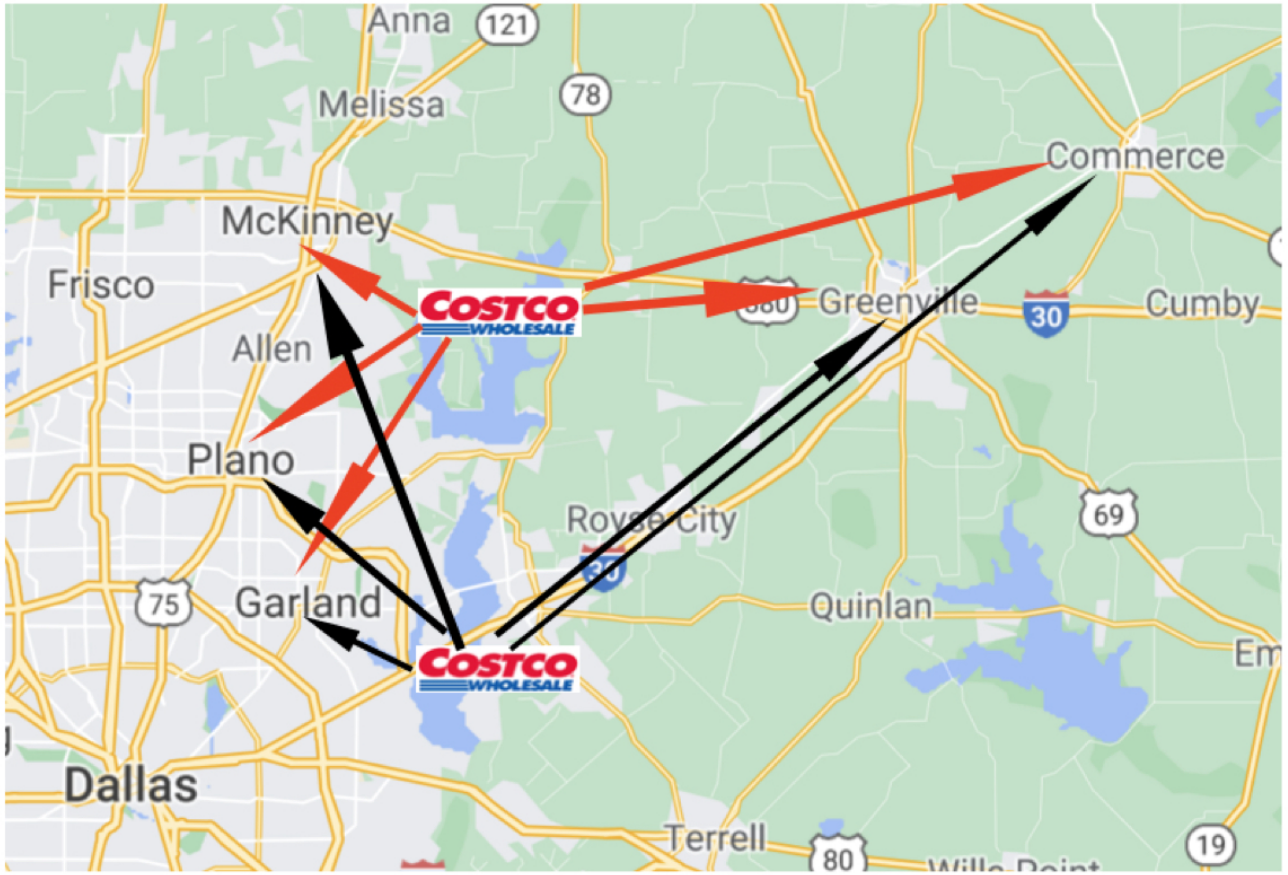
Topics of this lecture:

- k-center, k-median, k-means
- kernel, spectral clustering $k(\mathbf{x}, \mathbf{x}') \to |\mathbf{x} - \mathbf{x}'|^2$
- sparsity constraint, LASSO $\lambda \|\mathbf{w}\|_1$
- Huber loss, robustness "ℓ-2"+"ℓ-1"
- Gaussian mixing model
- EM algorithm, lower-bound function $\ln p(\mathbf{X}|\mathbf{w}) = \mathcal{L}(q, \mathbf{w}) + \mathrm{KL}(q\|p)$

Supermarket locations

$$d(\mathbf{x}^{(i)}, \vec{\mu}_k) = |\cdots|$$

2 Costcos considering
McKinney
Greenville
Commerce
Plano
Garland



k-median

$$\Phi_{\text{k-median}}(\mathcal{C}) = \min_{k=1}^{K} \sum_{\mathbf{x}^{(i)} \in C_k} d(\mathbf{x}^{(i)}, \vec{\mu}_k)$$

K=2, I=1~5

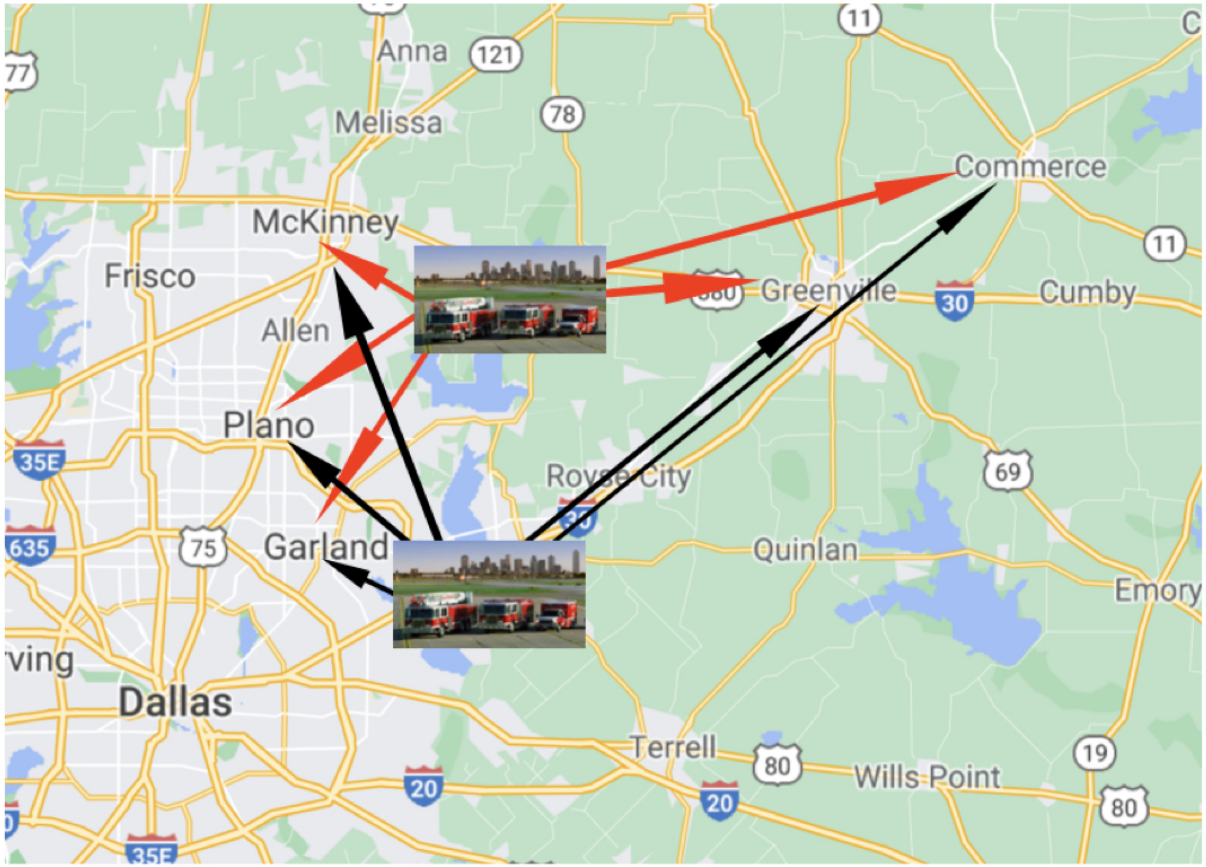
Fire-house locations

2 fire-houses considering
McKinney
Greenville
Commerce
Plano
Garland

k-center

$$\Phi_{\text{k-center}}(\mathcal{C}) = \min_{k=1}^{K} \max_{\mathbf{x}^{(i)} \in C_k} d(\mathbf{x}^{(i)}, \vec{\mu}_k)$$

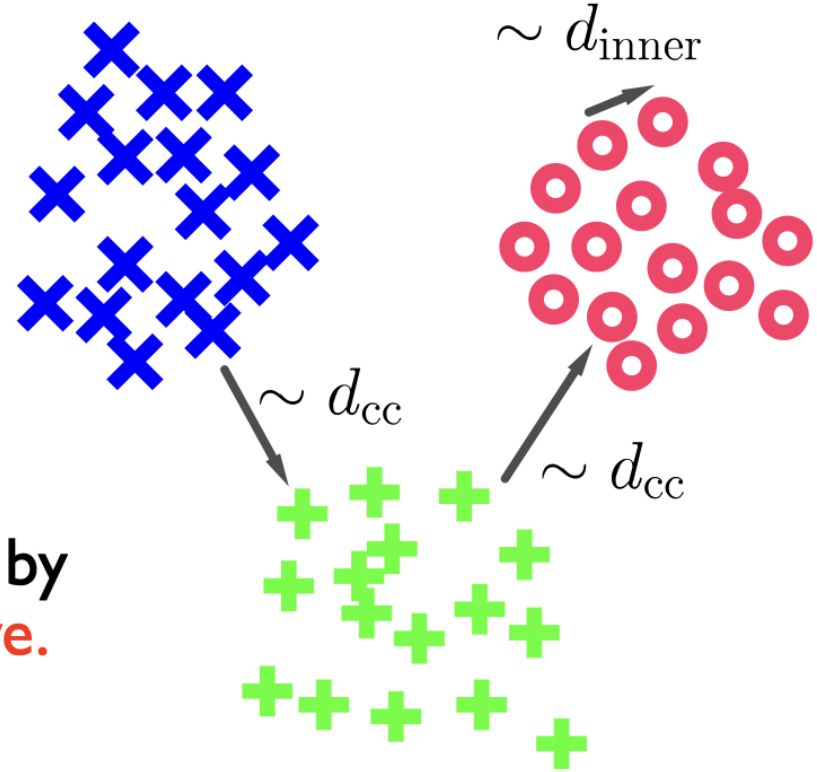
K=2, I=1~5



k-means

● outlier ?

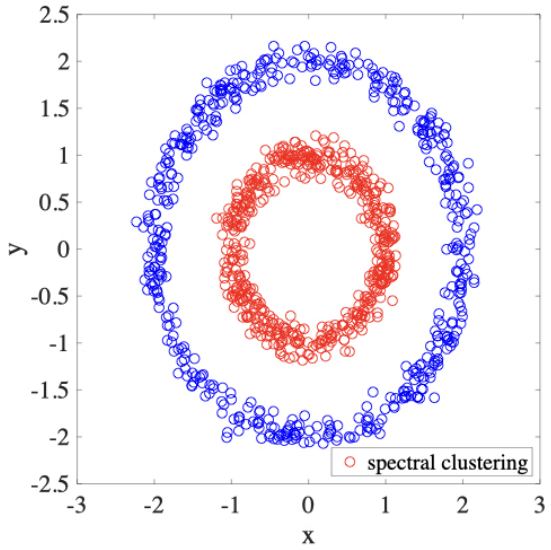
$$\Phi_{k\text{-means}}(\mathcal{C}) = \min_{k=1}^{K} \sum_{\mathbf{x}^{(i)} \in C_k} d^2(\mathbf{x}^{(i)}, \vec{\mu}_k)$$



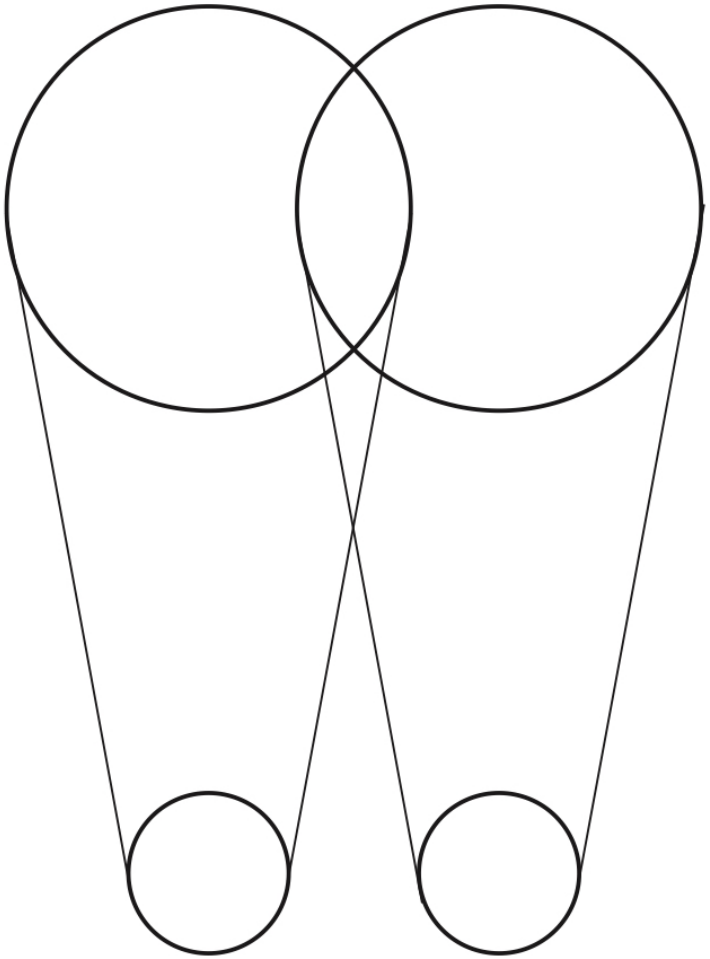
Implications: large value of distances is expected to play important role, consequently the scheme will be affected by strange points, namely it is outlier-sensitive.

Another motivation for clustering

the effective dimension of the data is low than its apparent representation, in these situations, it is better to project the data to the main/leading space or dimensions, the mathematical theory of it is the singular-value-decomposition (SVD) or the principal component analysis (PCA)



$re^{i\phi}$



Algorithm for k-means

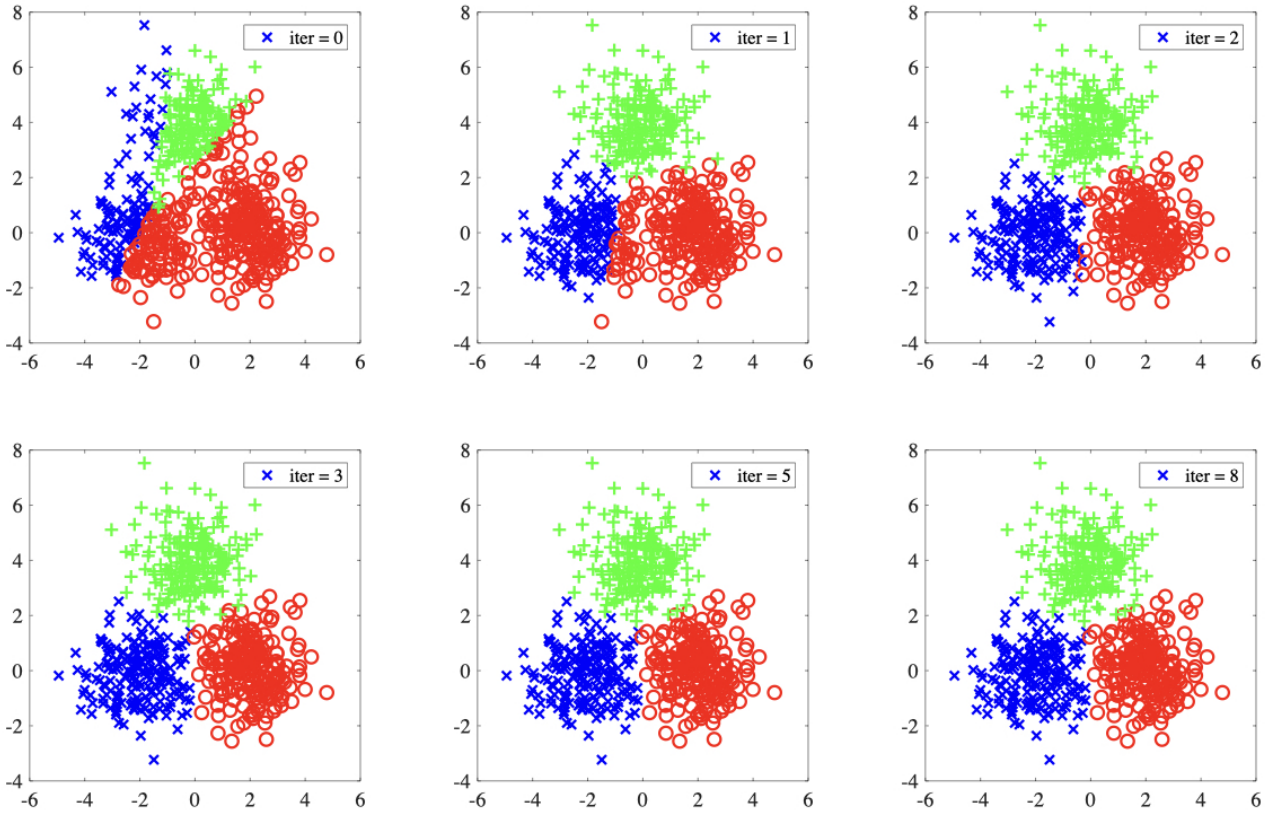
responsibility $r_k^{(i)} = 1$ if the sample $\mathbf{x}^{(i)}$ is in the kth center

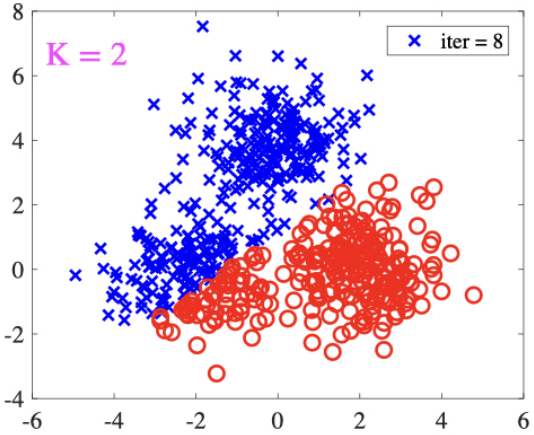
$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} r_k^{(i)} \|\mathbf{x}^{(i)} - \vec{\mu}_k\|^2$$

$$(r_k^{(i)}, \vec{\mu}_k)^* \leftarrow \text{argmin}_{r_k^{(i)}, \vec{\mu}_k} J$$

$r_k^{(i)} = 1$, if $k = \text{argmin}_j \|\mathbf{x}^{(i)} - \vec{\mu}_j\|^2$, $\vec{\mu}_k = \frac{\sum_{i=1}^{m} r_k^{(i)} \mathbf{x}^{(i)}}{\sum_{i=1}^{m} r_k^{(i)}}$ $\quad m_k = \sum_{i=1}^{m} r_k^{(i)}$

Example: K=3 or K=2





you can also use K=4

Sensitive to initialization (local optimal)

Lloyd problem/algorithm

initial centers {0,5,10}

when k is a part of the input or may be a function of m, the clustering optimization problems are generally NP-hard

Spectral clustering

$$\|\mathbf{x}^{(i)} - \vec{\mu}_k\|^2 = \left\| \mathbf{x}^{(i)} - \frac{1}{m_k} \sum_{j, \mathbf{x}^{(j)} \in C_k} \mathbf{x}^{(j)} \right\|^2 = \overbrace{\langle \mathbf{x}^{(i)} | \mathbf{x}^{(i)} \rangle}^{k(\mathbf{x}^{(i)}, \mathbf{x}^{(i)})} - \frac{2}{m_k} \sum_{j, \mathbf{x}^{(j)} \in C_k} \langle \mathbf{x}^{(i)} | \mathbf{x}^{(j)} \rangle + \frac{1}{m_k^2} \sum_{j, j', \mathbf{x}^{(j)}, \mathbf{x}^{(j')} \in C_k} \langle \mathbf{x}^{(j)} | \mathbf{x}^{(j')} \rangle$$

$$\text{independent of } k$$

kernel: $k(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x} | \mathbf{x}' \rangle$

$$\{C_1, C_2, \cdots, C_k\} \leftarrow \text{argmin}_{k \in \{1, \cdots, K\}} \left[-\frac{2}{m_k} \sum_{j, \mathbf{x}^{(j)} \in C_k} k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) + \frac{1}{m_k^2} \sum_{j, j', \mathbf{x}^{(j)}, \mathbf{x}^{(j')} \in C_k} k(\mathbf{x}^{(j)}, \mathbf{x}^{(j')}) \right]$$

Warm-up: Lagrange multiplier

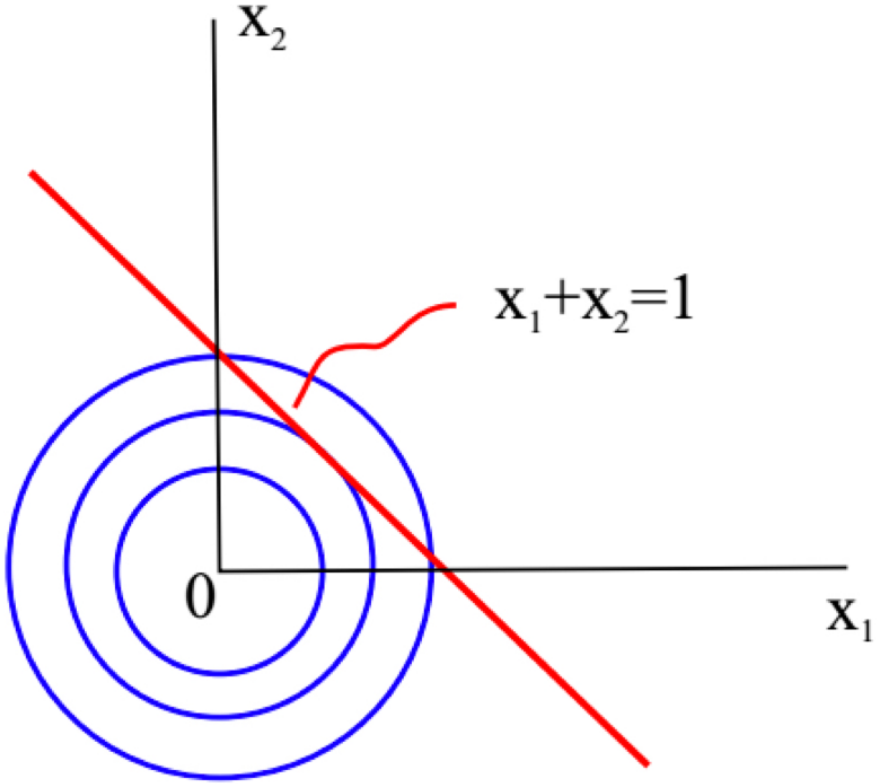
aim: $\min_{x_1,x_2} f(x_1,x_2),\ f(x_1,x_2) = 1 - x_1^2 - x_2^2$

constraint: $g(x_1,x_2) = 1 - x_1 - x_2 = 0$

Lagrange function:

$$L(\mathbf{x},\lambda) = f(\mathbf{x}) - \lambda g(\mathbf{x})$$

$$\boxed{\frac{\partial L}{\partial \mathbf{x}} = \vec{0},\ \frac{\partial L}{\partial \lambda} = 0}$$



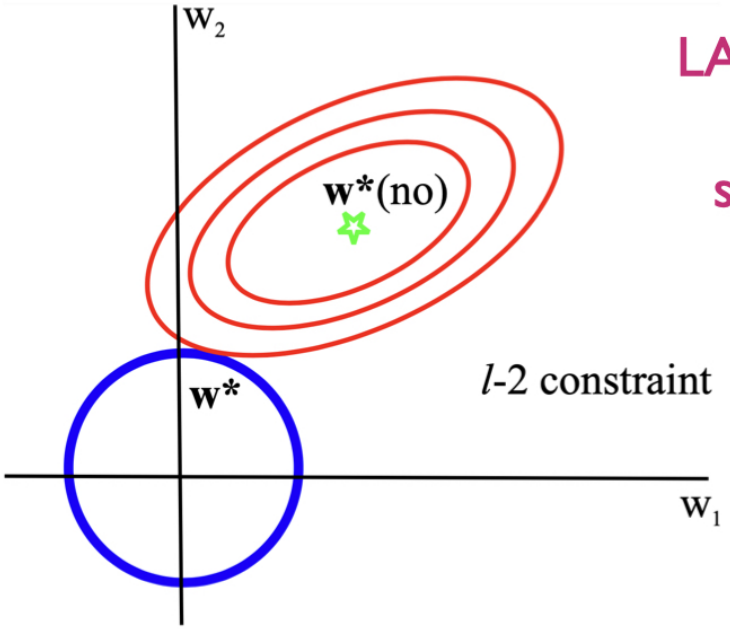
Sparsity: geometrical explanation

ℓ-2 loss + ℓ-2 constraint:

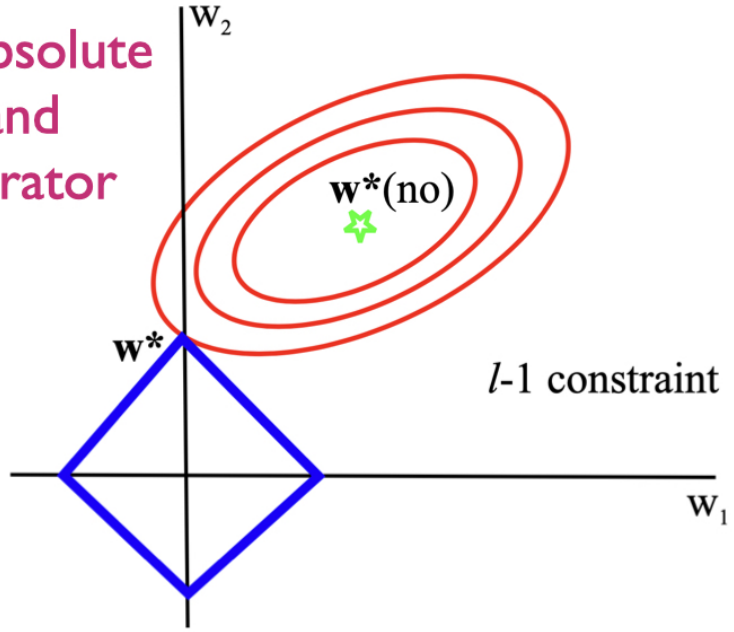
$$\mathbf{w}^* = \text{argmin}_{\mathbf{w}} \frac{1}{2} \left\| \vec{\Phi}\mathbf{w} - \mathbf{y} \right\|_2^2 + \lambda \|\mathbf{w}\|_2^2$$

ℓ-2 loss + ℓ-1 constraint:

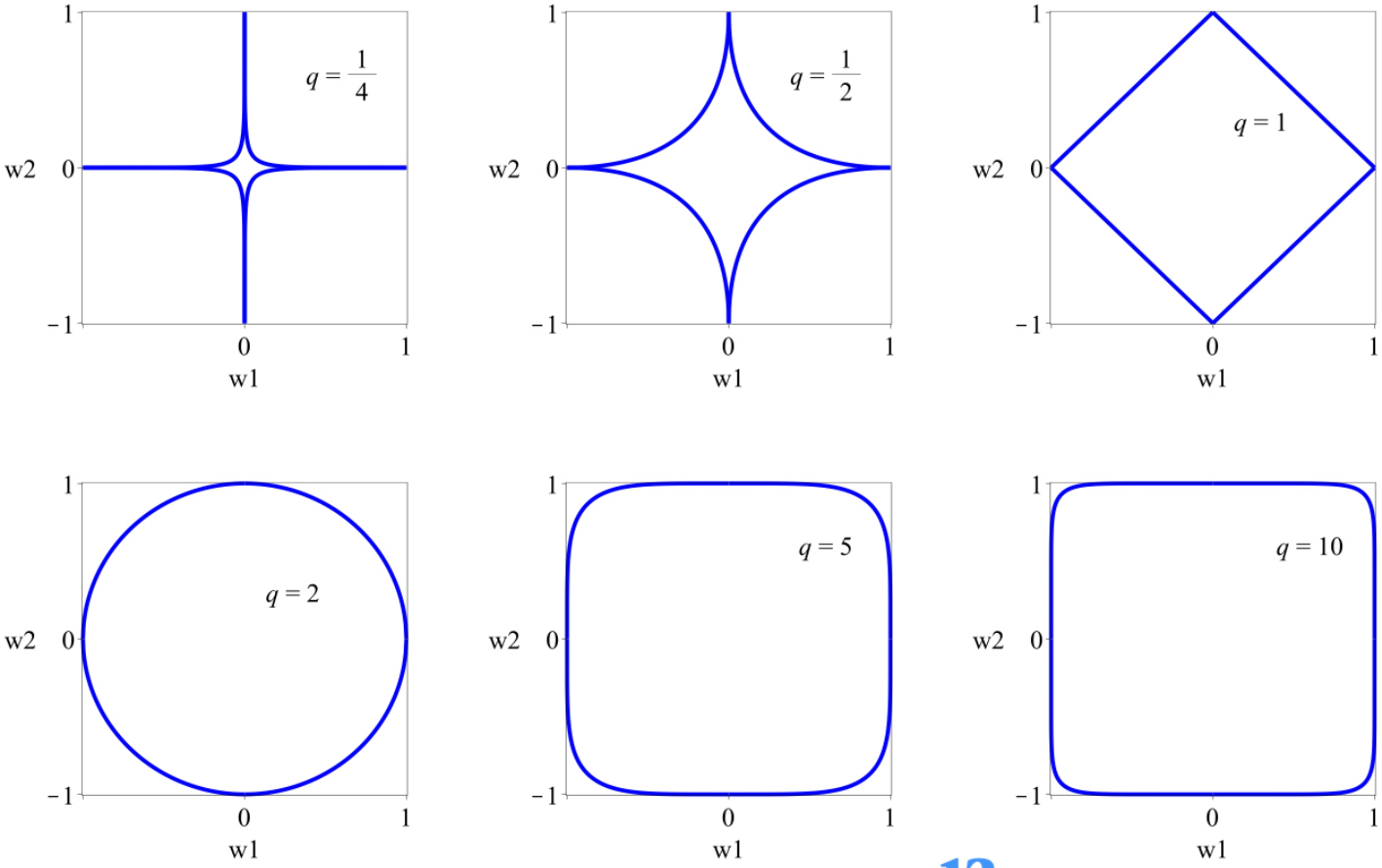
$$\mathbf{w}^* = \text{argmin}_{\mathbf{w}} \frac{1}{2} \left\| \vec{\Phi}\mathbf{w} - \mathbf{y} \right\|_2^2 + \lambda \|\mathbf{w}\|_1$$



LASSO: least absolute shrinkage and selection operator



q-ball for constraints

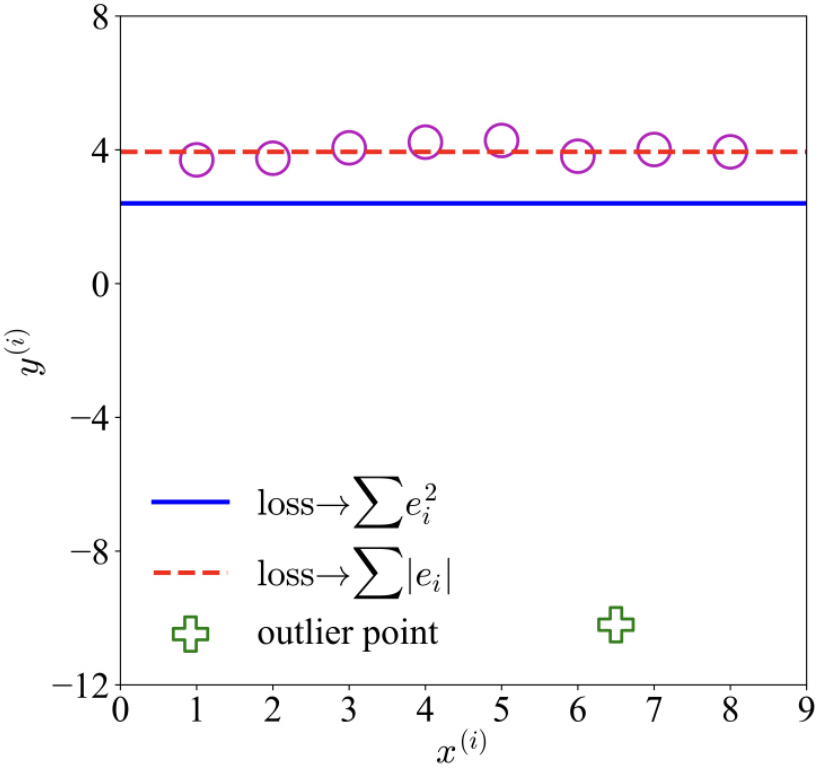


elastic net

$$\alpha\|\mathbf{w}\|_2^2 + \beta\|\mathbf{w}\|_1$$

Robustness: concept and example

$x^{(i)}$	1	2	3	4	5	6	7	8	6.5
$y^{(i)}$	3.70	3.75	4.06	4.23	4.28	3.81	4.01	3.94	−10.20



ℓ-2 loss: $\langle y \rangle$

ℓ-1 loss: median$(y^{(i)})$

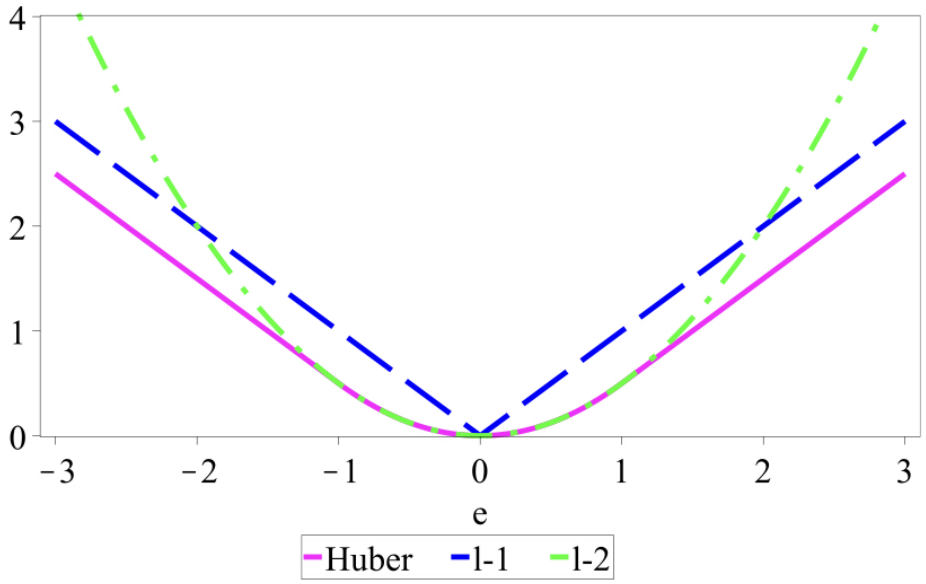
Least-squares is affected by outliers!

Ex.: expliain why?

Huber loss

$$e_i = f_{\mathbf{w}}(x^{(i)}) - y^{(i)}$$

$$\chi^{\mathrm{H}}(e) = \begin{cases} 2^{-1}e^2, & |e| \le \eta \\ \eta|e| - 2^{-1}\eta^2, & |e| > \eta \end{cases}$$



Optimization with Huber loss

(1) Initialized $\mathbf{w}$ via least-squares:

$$\mathbf{w} \leftarrow (\vec{\Phi}^\top \vec{\Phi})^{-1} \vec{\Phi}^\top \mathbf{y}$$

(2) Calculate weight $\vec{\Theta}$ using current $\mathbf{w}$:

$$\vec{\Theta} = \mathrm{diag}(\theta^{(1)}, \cdots, \theta^{(m)})$$

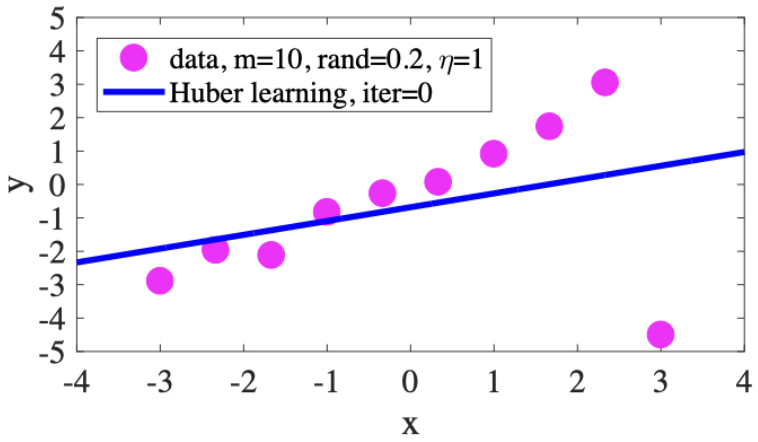
$$\text{with } \theta^{(i)} = \begin{cases} 1, & |e_i| \le \eta \\ \eta/|e_i|, & |e_i| > \eta \end{cases}$$

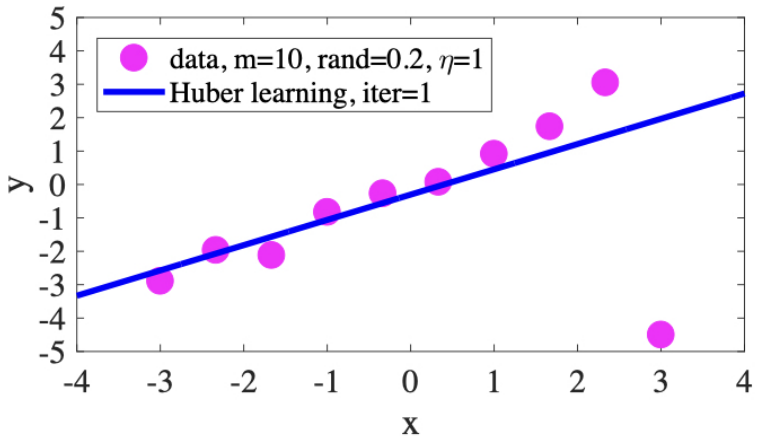
(3) Recalculate the learning parameter:

$$\mathbf{w} \leftarrow (\vec{\Phi}^\top \vec{\Theta} \vec{\Phi})^{-1} \vec{\Phi}^\top \vec{\Theta} \mathbf{y}$$

Ex.: what's the effect of Huber loss + ℓ-1 constrain?

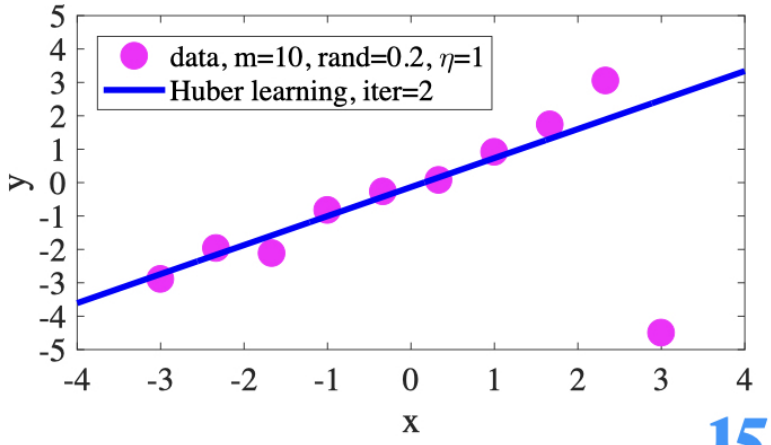
Example on using Huber loss

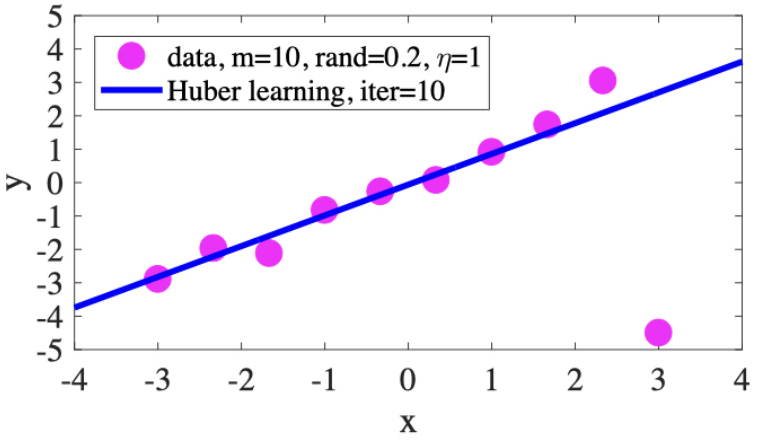




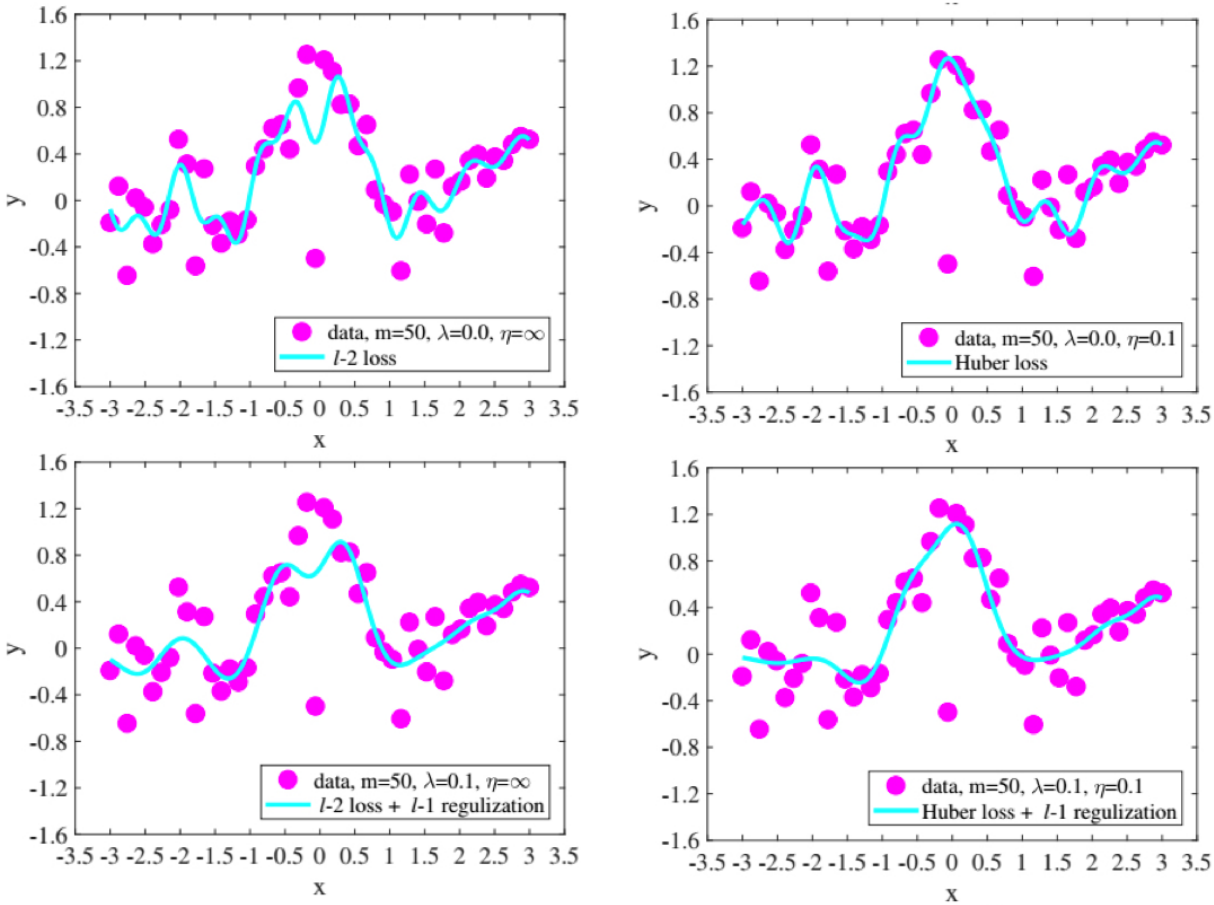
$y^{(i)} = x^{(i)} + \text{rand} \times \delta$

outlier: 1





Huber loss + sparse constraint



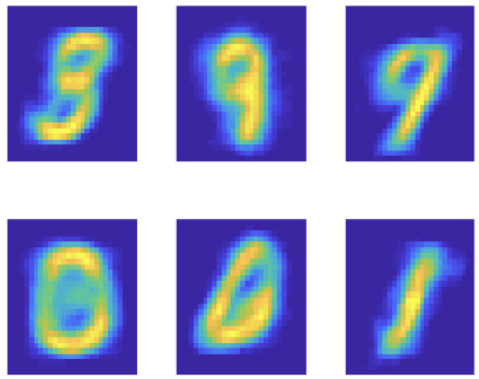
$$f_{\mathbf{w}}(\boldsymbol{x}) = \sum_{j=1}^{m} w_j \exp(-\|\mathbf{x} - \mathbf{x}^{(j)}\|^2 / 2\nu^2)$$

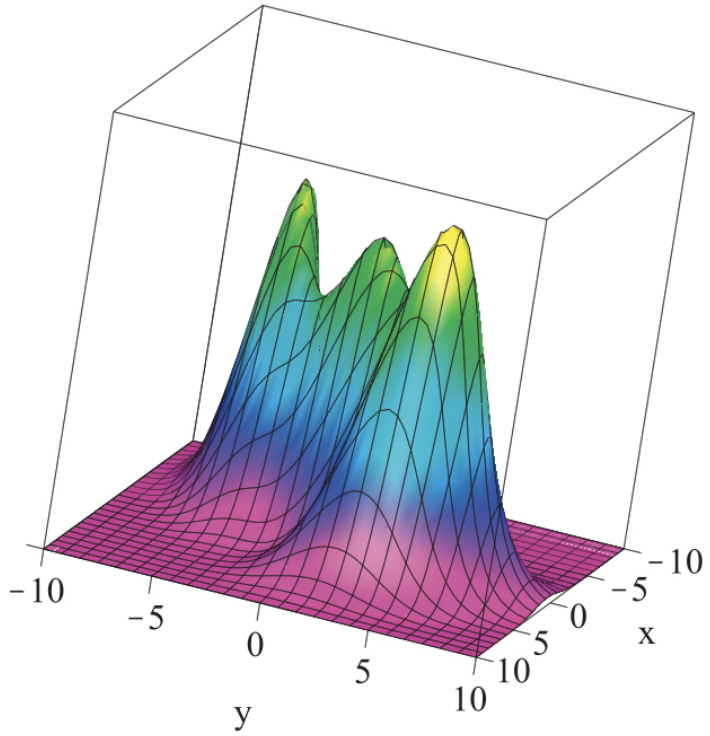
- sparsity improves outlier
- Huber loss and sparsity improves the overall smoothness of the curve

Ex.: now write down the algorithm flow of Huber+sparsity

Gaussian mixing model (GMM)

$$p(\mathbf{x}) = \sum_{k=1}^{K} \zeta_k \mathcal{N}(\mathbf{x}|\vec{\mu}_k, \vec{\Sigma}_k), \sum_{k=1}^{K} \zeta_k = 1$$





$$\gamma(\mathbf{z}_k) = \frac{\zeta_k \mathcal{N}(\mathbf{x}|\vec{\mu}_k, \vec{\Sigma}_k)}{\sum_{k'} \zeta_{k'} \mathcal{N}(\mathbf{x}|\vec{\mu}_{k'}, \vec{\Sigma}_{k'})}$$

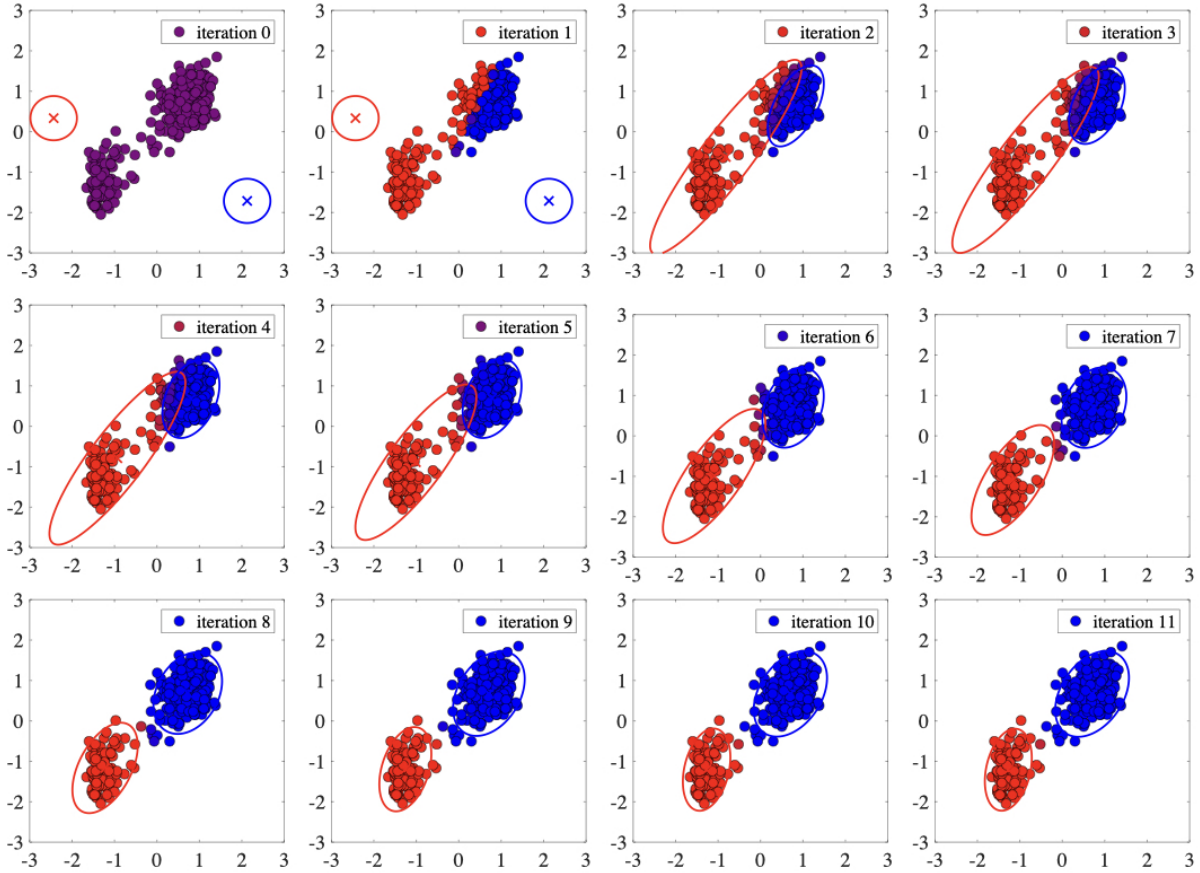
$$\vec{\mu}_k = \frac{1}{m_k} \sum_{i=1}^{m} \gamma(\mathbf{z}_k^{(i)}) \mathbf{x}^{(i)}, \; m_k = \sum_{i=1}^{m} \gamma(\mathbf{z}_k^{(i)})$$

$$\vec{\Sigma}_k = \frac{1}{m_k} \sum_{i=1}^{m} \gamma(\mathbf{z}_k^{(i)}) \left(\mathbf{x}^{(i)} - \vec{\mu}_k\right) \left(\mathbf{x}^{(i)} - \vec{\mu}_k\right)^{\top}$$

$$\zeta_k = m_k / m$$

Ex.: show that the GMM approaches to k-means if the width->0

Example using GMM



auto-encoder; EM algorithm, …

lower bound function: $\mathcal{L}(q, \mathbf{w})$
functional of distribution q
function of parameter $\mathbf{w}$

$$\ln p(\mathbf{X}|\mathbf{w}) = \mathcal{L}(q, \mathbf{w}) + \underbrace{\mathrm{KL}(q\|p)}_{\text{non-negative}}$$