Clustering

Task: Find clusters / subgroups in a dataset $S = \{x_1, \dots, x_n\} \in \mathbb{R}^p$.

- ▷ Samples within subgroup similar / homogeneous
- ▷ Samples in different subgroups "distant" / heterogeneous from each other.

Ex:
- Customers of a company ⟶ grouping for targeted marketing.
- Biology: Find groups of genes.

Q:
- What notion of similarity?
- Precise definition?

"Unsupervised Learning": No subset of data with "correct" classification / grouping available.

Note: Fundamentally different from classification problems!

K-means clustering: [Steinhaus '56, Lloyd '57]

Given n points $x_1,\dots,x_n \in \mathbb{R}^p$, find K centroids $c_1,\dots,c_K \in \mathbb{R}^p$ and a partition $\Gamma_1 \cup \Gamma_2 \cup \dots \cup \Gamma_K = \{1,\dots,n\}$ ($\Gamma_i \cap \Gamma_j = \emptyset$ f $i \neq j$) such that

$$F\left(\{c_i\}_{i=1}^{k}, (\Gamma_i)_{i=1}^{k}\right) := \sum_{j=1}^{k} \sum_{i \in \Gamma_j} d(x_i, c_j)$$

"k-means objective" (*)

is minimized, where $d: \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ is a "distance" function

Ex: 1. $d(x,y) = \|x-y\|_2^2$ "squared Euclidean" ← by default

2. $d(x,y) = \|x-y\|_1$. "k-median".

Observation: K-means with is NP-hard for $k \geq 2$. [Drineas et al. '04].

Lloyd's algorithm (often called "k-means"):

Input: $\{x_i\}_{i=1}^n \in \mathbb{R}^p$, desired nr. of clusters k

1. Initialize $c_1, \dots, c_k \in \mathbb{R}^p$.

a) randomly among the $\{x_i\}_{i=1}^n$

or

b) k-means++: fancier

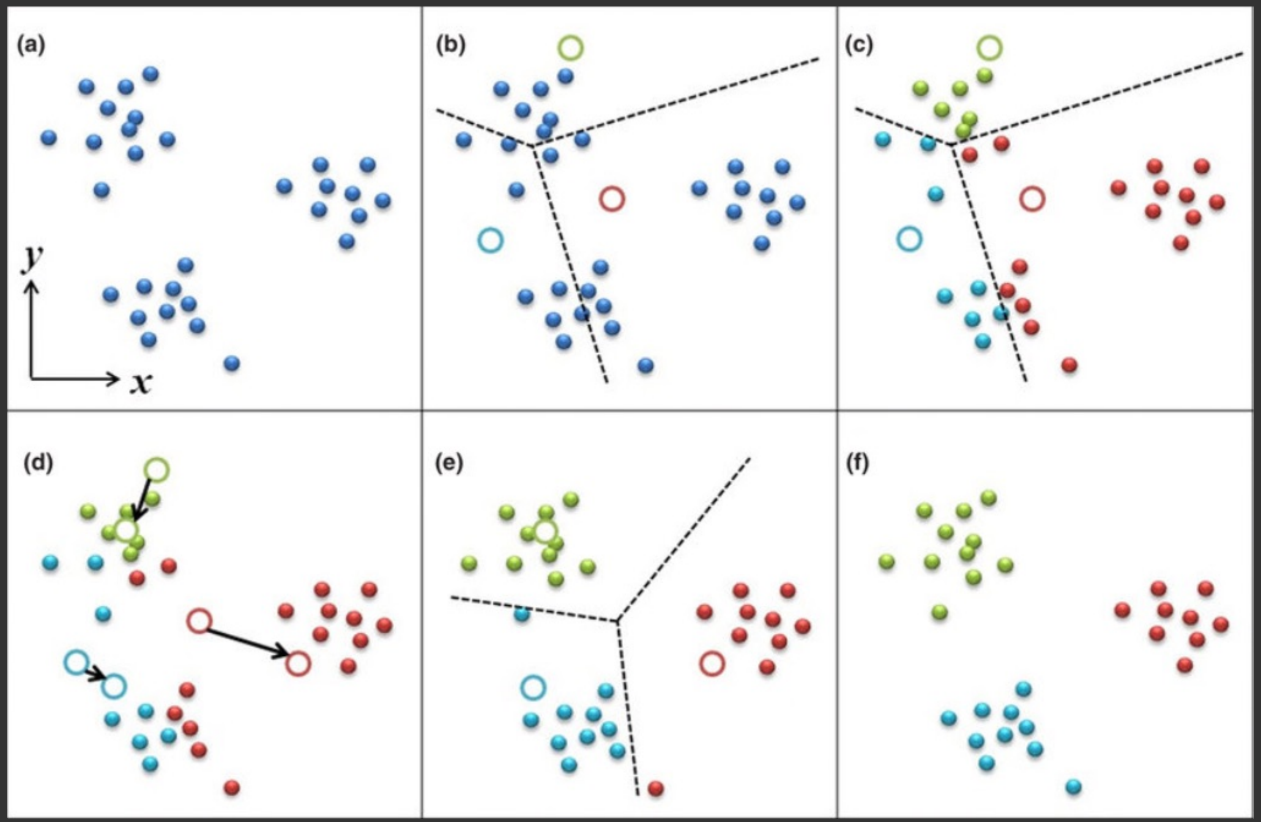
Repeat until convergence:

2. $\forall i = 1, \dots, n$: Assign $x_i \in \Gamma_j$ if c_j is closest centroid to x_i among $\{c_\ell\}_{\ell=1}^k$

3. Update $\forall j = 1, \dots, k$:

$$c_j = \underset{c \in \mathbb{R}^k}{\arg\min} \sum_{i \in \Gamma_j} d(x_i, c) = \frac{1}{|\Gamma_j|} \sum_{i \in \Gamma_j} x_i$$

↖ if $d(\cdot,\cdot) = \|\cdot - \cdot\|_2^2$



▷ Finds local optimum of (*)

▷ Works well for "convex" clusters

To consider:

- ▷ How to choose nr. of clusters k?
- ▷ Which distance to choose (geometry of underlying space..)?
- ▷ Initialization: If prior knowledge available, → might be better than random.
- ▷ Needs a lot of pairwise distances. If $n >> 10^5$ or so, slow → "Minibatch KMeans"

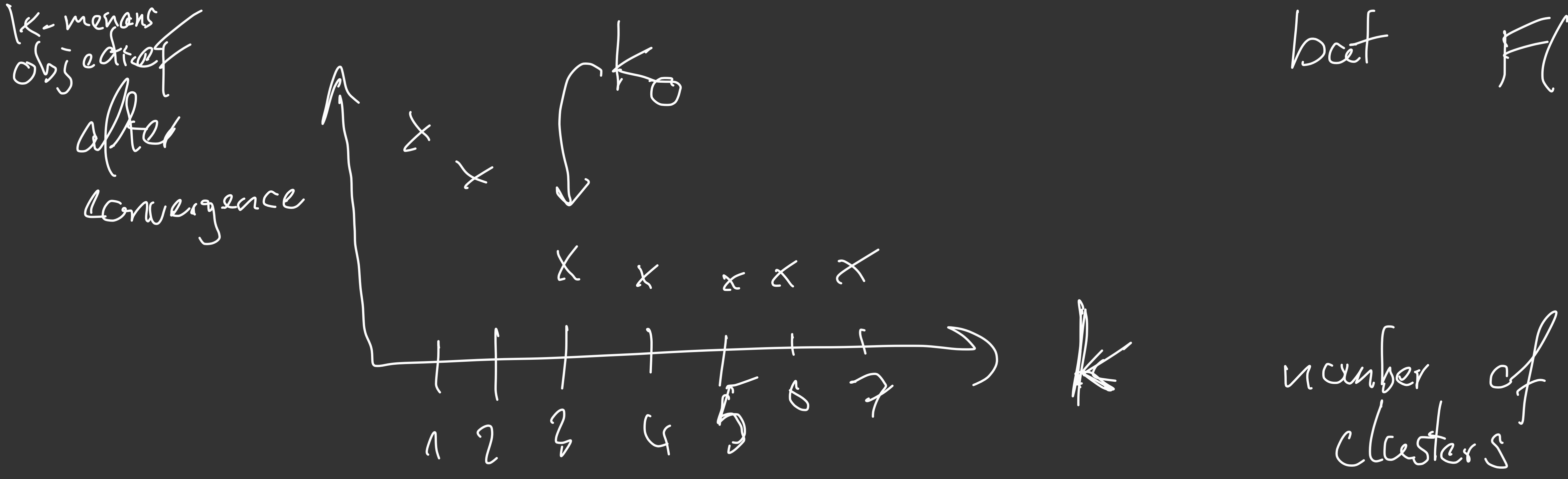
Other clustering methods:

- Spectral Clustering: Based on laplacian of similarity graph.
- Hierarchical Clustering: Creates trees

How to choose parameter "k" in pratice?

One option: Run Lloyd's algorithm for $k = 1, 2, \dots$ until convergence and then find some k_0 s.t. $F(k_0) << F(k_0 - 1)$,

but $F(k_0 + \ell) \simeq F(k_0)$ for many $\ell = 1, 2, \dots$,



"Elbow plot"