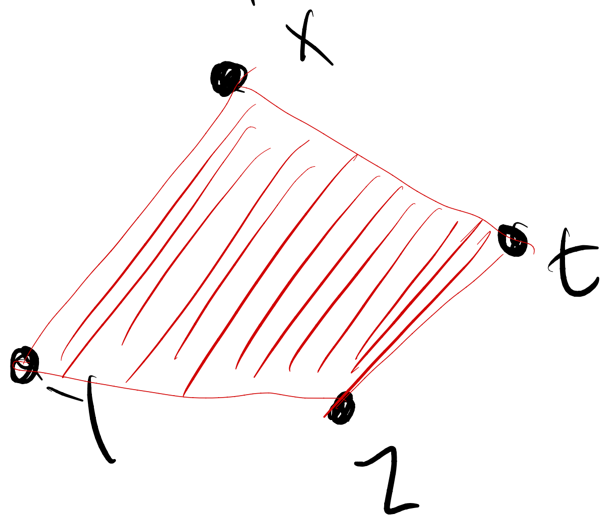
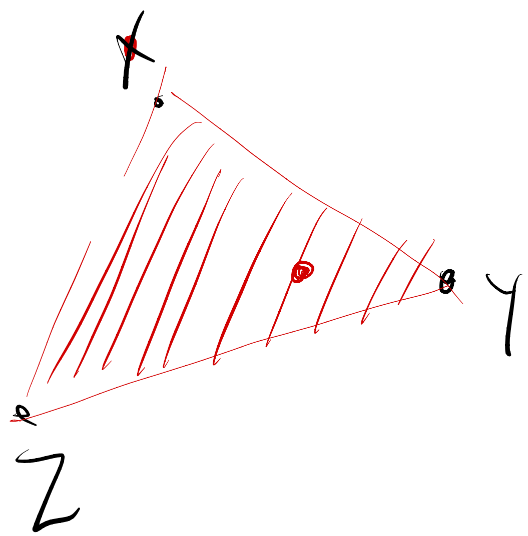
Clustering - KMEANS

convexity - closure = $\{a \mid a = c_1 x + c_2 y + c_3 z + c_4 t\}$

$c_1, c_2, c_3, c_4 \in \mathbb{R}^+$

$c_1 + c_2 + c_3 + c_4 = 1$

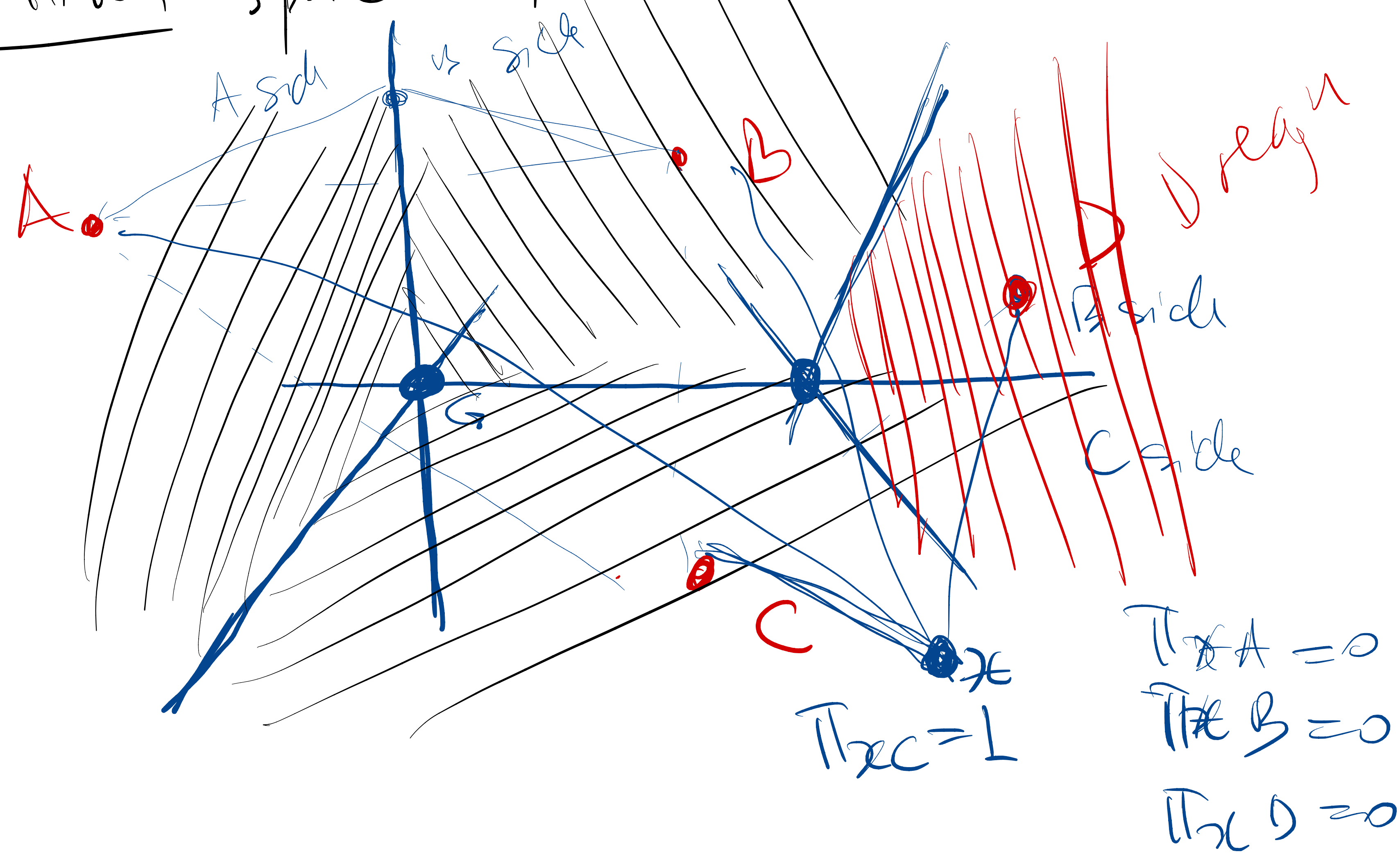




Space Partition by distance - idea

A B C centroids

<u>partition</u> space by min-dist to A, B, C.



x_i = datapoint i $\quad i = 1:N \quad K = \#$ of clusters

μ_k = centroid for k-th group/ cluster $\quad k = 1:K$

π_{ik} = membership indicator $= \begin{cases} 1 & \text{if } x_i \to \text{cluster } k \\ 0 & \text{otherwise} \end{cases}$

Decide Membership / Given centroids μ_k

E step

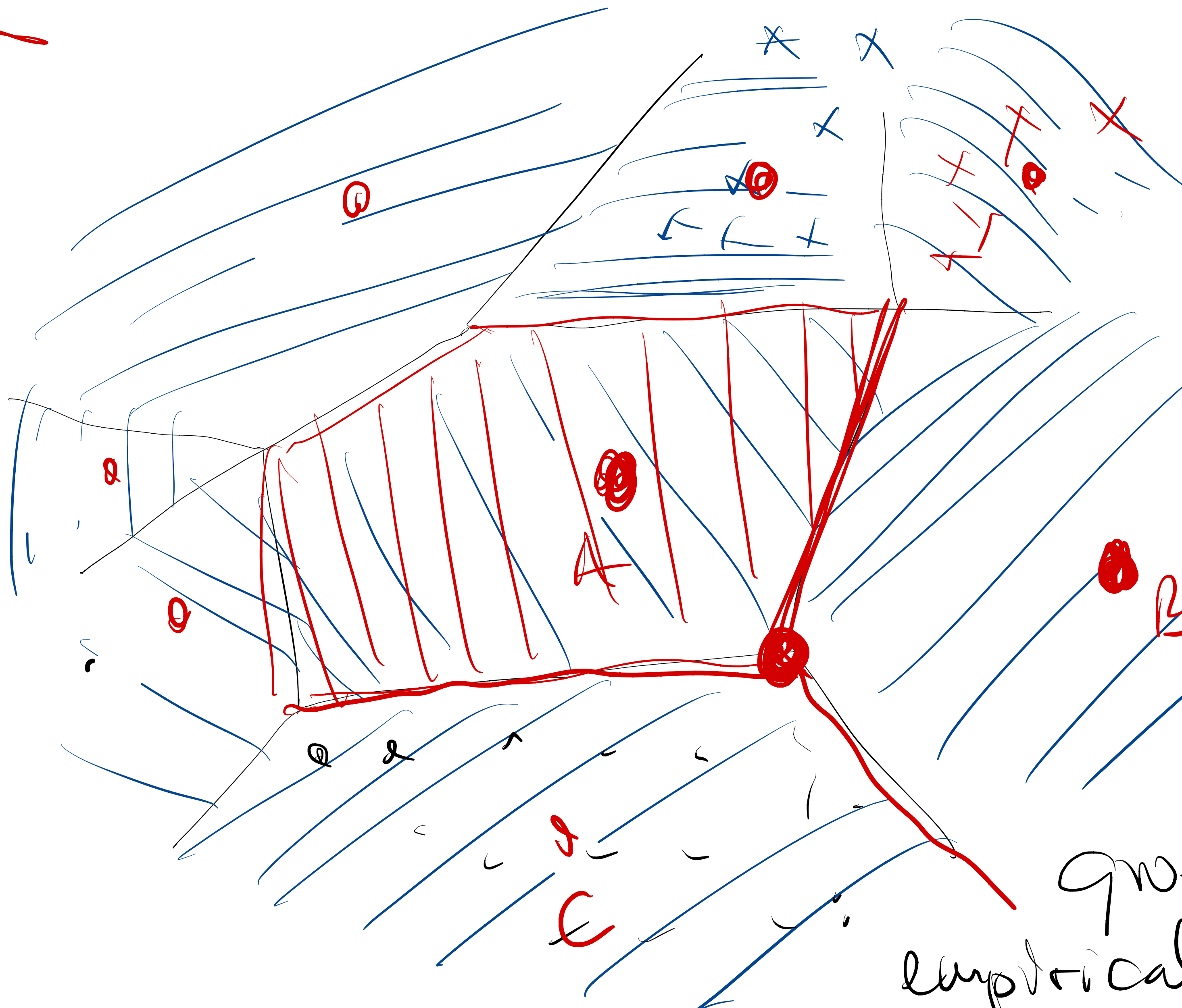
$\boxed{\text{for } x_i}$

$$\text{cluster}_k = \arg\min_k \{ \text{dist}(x_i, \mu_k) \}$$

$\pi_{ik} = 1$ for that argmin, 0 for the others

Decide Centroids / Given Membership This

M step



theoretical
centers for
complete
space partition

practice:
groups of datapoints
empirical centroids

μ_k = avg of datapoints in cluster k

k-fixed $= \dfrac{\sum_{x \in \text{cluster } k} x}{\# x \text{ in cluster } k}$ → sum of vectors

$= \dfrac{\sum_{i=1}^{N} \pi_{ik} \cdot x_i}{\sum_{i=1}^{N} \pi_{ik}}$

optimal for $dist(x, \mu) = \|x - \mu\|^2$

Medium (Global)

K Means: improve iteratively E step / M step until convergence

dist$_2$ Euc

fix k

$SSE = \text{obj} \; \min_i \left\{ \sum_i \pi_{ik} \cdot dist(x_i, \mu_k) \right\} + \cancel{\ldots}$

proof

- easy : E step ⇒ optimal π_{ik} | given μ_k
- easy M step ⇒ optimal μ_k | given π_{ik}

- not easy global objective

$$\min \left\{ \sum_{i,j,k} (\pi_{ik} \cdot \pi_{jk}) \, \text{dist}(x_i, x_j) \right)$$

points in same cluster (k) have small dist on avg

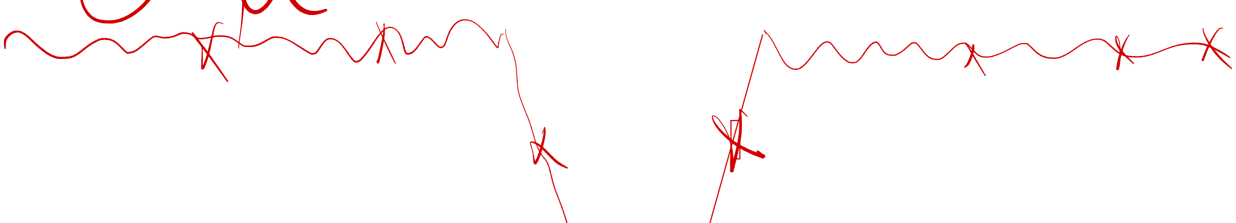
OBS :

OBJ SSE $\sum_i \sum_k \pi_{ik} \|x_i - \mu_k\|^2$

$$= \sum_k \left\{ \sum_{\substack{i \\ \pi_{ik}=1}} \|x_i - \mu_k\|^2 \right\}$$

MIN $= \sum_k \left\{ \sum_i \pi_{ik} \|x_i - \mu_k\|^2 \right\}$

$$\frac{\partial \text{OBJ}}{\partial \mu} = 0 \xRightarrow{\text{exercise}} \mu_k = \frac{\sum \pi_{ik} \cdot x_i}{\sum \pi_{ik}}$$



Evaluation Clustering

idea ① All pairs $\binom{N}{2}$

All pairs same cluster

$$\sum_{i,j} sim(x_i, x_j)$$

$\pi_{ik} = \pi_{jk} \ \forall k$

big

All pairs diff cluster

$$\sum_{i,j} sim(i,j)$$

$\pi_{ik} \neq \pi_{jk} \ \exists k$

small

idea 2 — supervised (labels/tags correlated with desired clusters)

BETTER EVAL

same label ⟹ same cluster

same cluster ⇒ same/similar labels

⇒ small label variance

* deal with granularity/specificity
  - clusters
  - tags/labels