FoDA L23

K-Means

Clustering

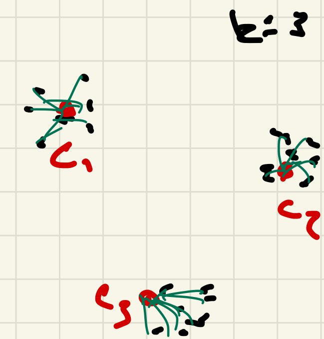
Assignment-Based Clustering

Input $X \subset \mathbb{R}^d$ $\quad X = \{x_1, x_2 \dots x_n\}$

distance $\quad d: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$

$$d(x,p) = \|x-p\|$$

Value k



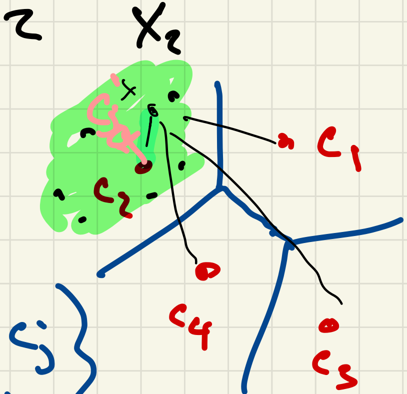
Goal Set of sites $S = \{c_1, \dots c_k\} \subset \mathbb{R}^d$

minimize $$f_{x \in X}\big(\underbrace{d(x, \phi_S(x))}_{\text{residual}}\big)$$

maps x to closest $c_i \in S$.

- $\sum_{x \in X} \|x - \phi_S(x)\|^2$ k-means
- $\sum_{x \in X} \|x - \phi_S(x)\|$ k-median
  k-medoid $S \subset X$
- $\max_{x \in X} \|x - \phi_S(x)\|$ k-center

$$\mathrm{Cost}(X, S) = \sum_{x \in X} \| x - \phi_S(x) \|^2$$



Lloyd's Algorithm

$X_j := \{ x \in X \mid \phi_S(x) = c_j \}$

sets

create

1a

0. Init: Choose k points $S \subset X$

   $S = \{c_1, \dots c_k\} \subset \mathbb{R}^d$

   sites

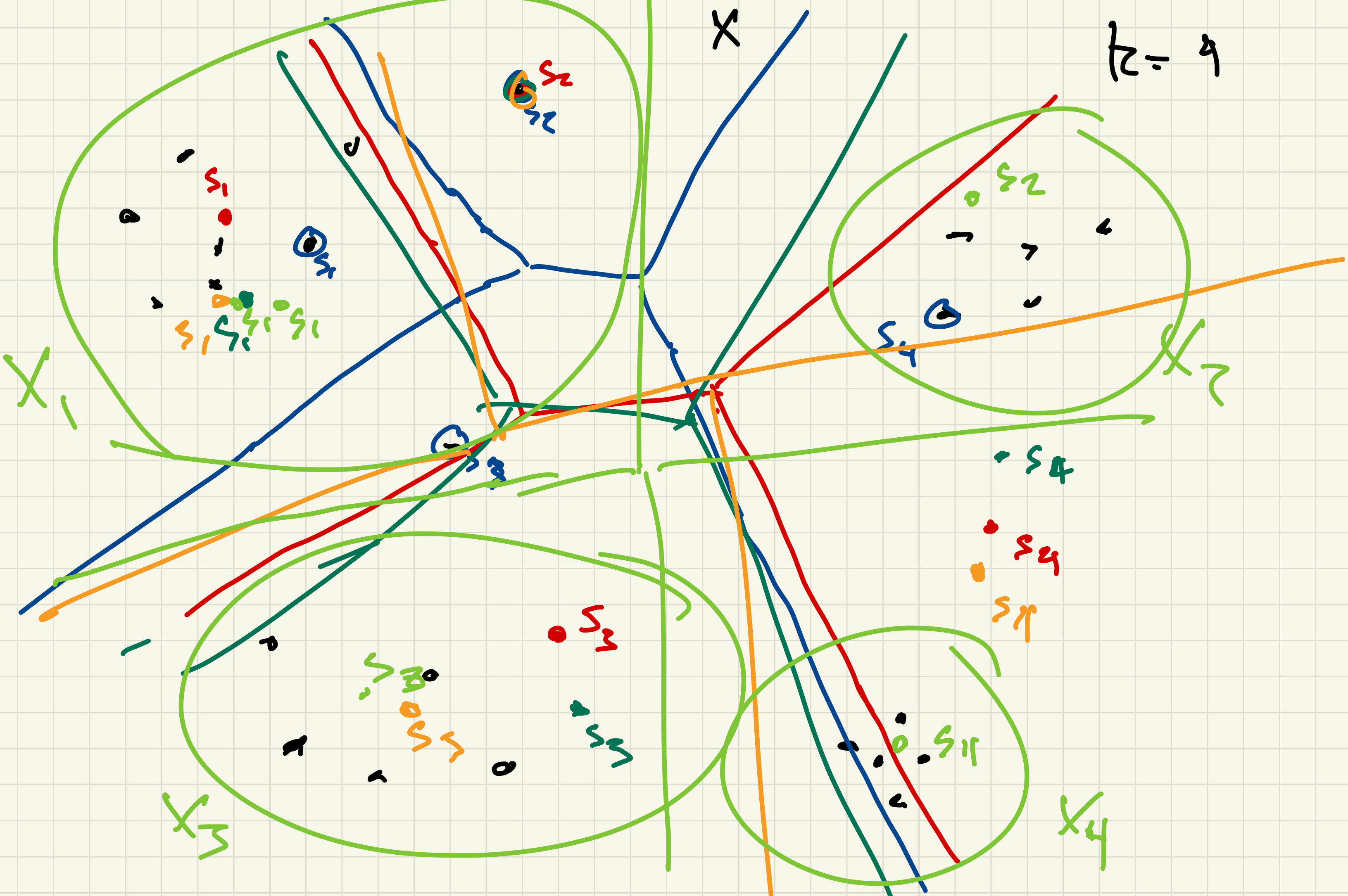
1. repeat

   1a for all $x \in X$: assign x to X_j s.t. $\phi_S(x) = c_j$ update sets

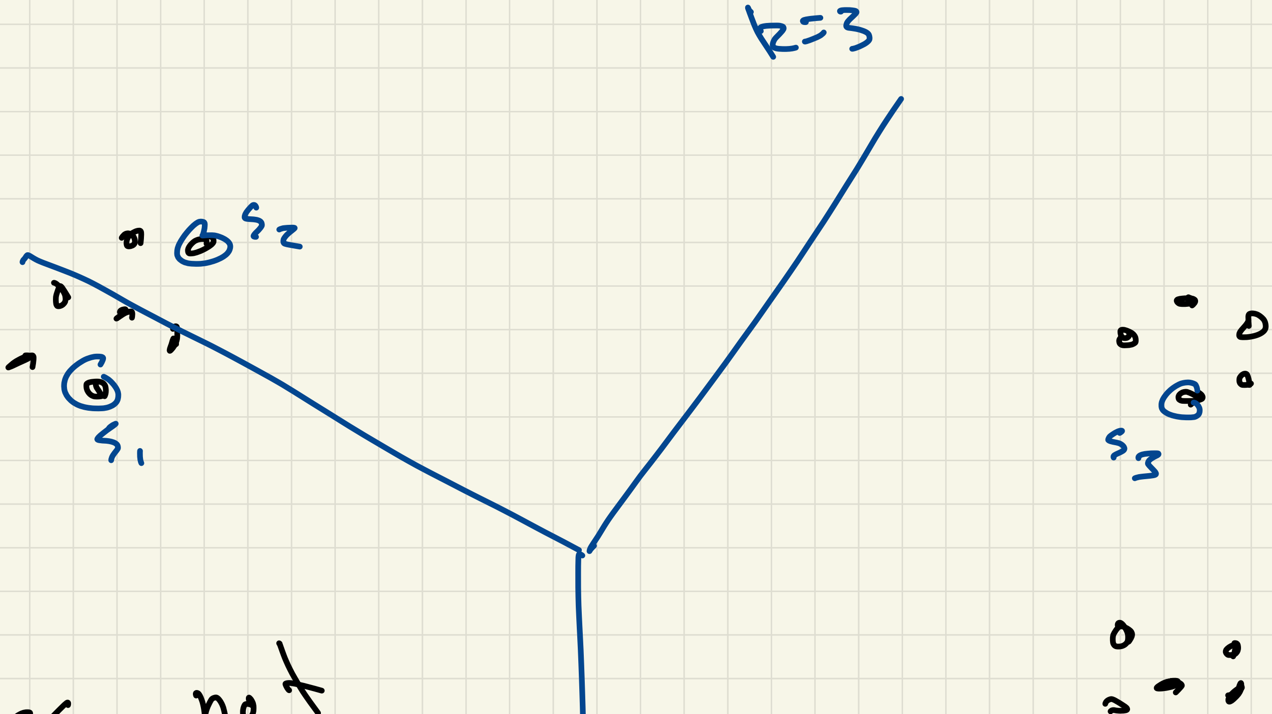
   1b for all $c_j \in S$: update $c_j = \frac{1}{|X_j|} \sum_{x \in X_j} x$ update sites

   until (R rounds or X_j unchanged)

2. Return S



X
k = 4
s1
s2
s3
s4
X1
X2
X3
X4

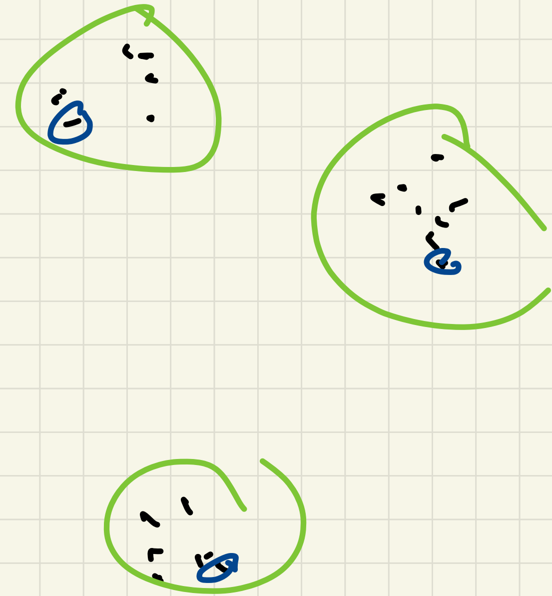


Does not always converge to optimal solution!

Initialization is important

Goal: 1 site in each true cluster.

option 1: plot, choose, optimize



option 2: ensure all dist
$\|c_j - c_{j'}\|$ large

Gonzalez alg.
k-means++

option 3: Random: choose $S \subseteq X$ at random

option 4: Use large $k' = 10 \cdot k$) → then merge.

Random Restarts

best-score $= \infty$

for $j = 1 \ldots T$ steps

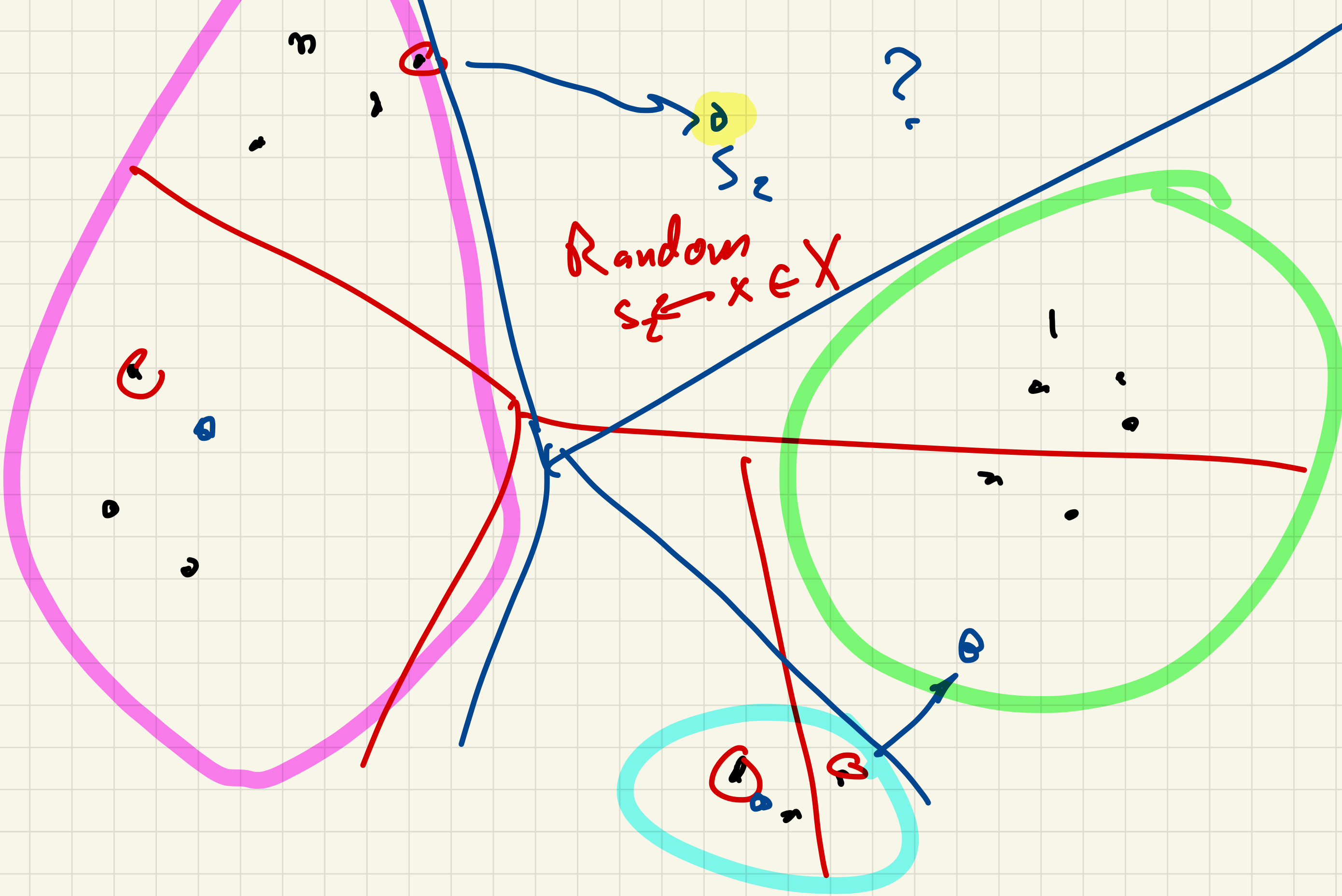
0. Random init S_j
1. Run Lloyds on (X, S_j)
2. if $\text{Cost}(X, S_j^*) < \text{best-score}$

   $S^* = S_j^*$

   $\text{best-score} = \text{Cost}(X, S_j^*)$

Return S^*

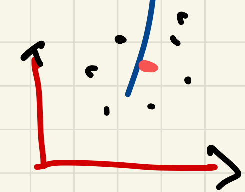
$$X_z = \{ x \in X \mid \phi_s(x) = s_z \}$$



Convergence Lloyd's Algo.

$$\text{Cost}(X, S) = \sum_{x \in X} \| x - \phi_S(x) \|^2 \qquad (1)$$

$$= \sum_{c_j \in S} \left(\sum_{x \in X_j} \| c_j - x \|^2 \right) \qquad (2)$$



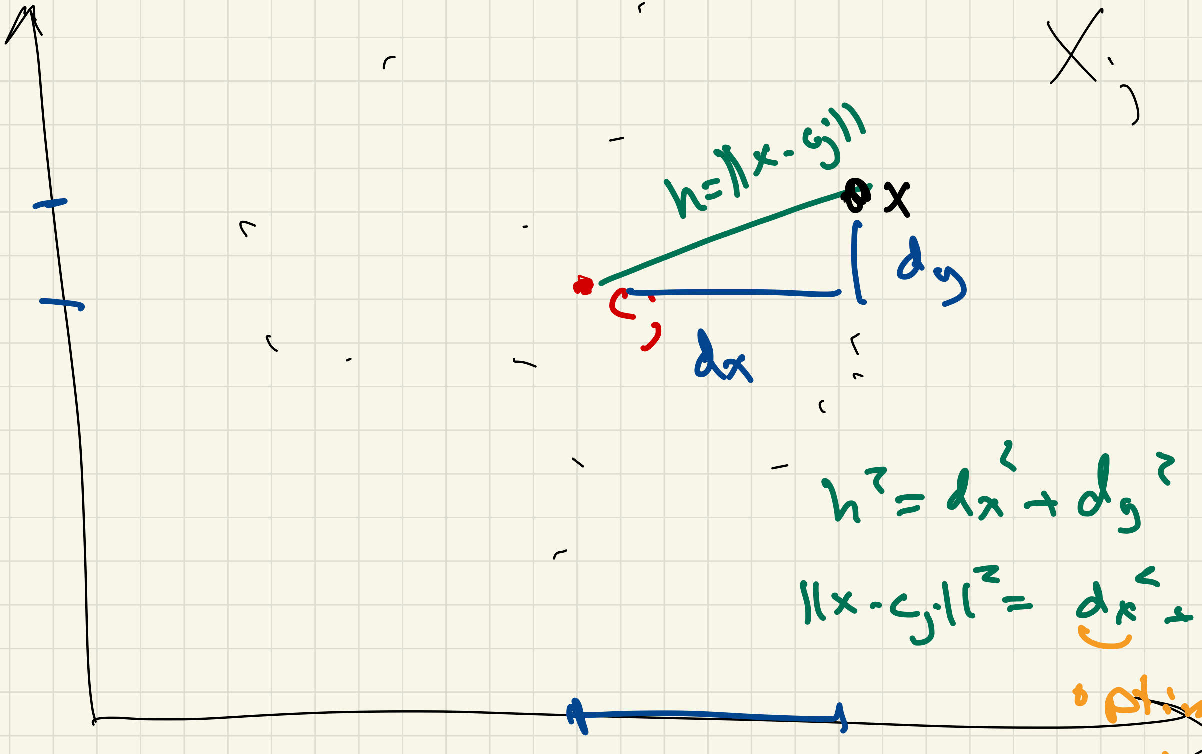
Lloyd's

1a optimize $X_j = \{ x \in X \mid \phi_S(x) = c_j \}$

proved in $d = 1$
average minimizer $\sum (\cdot)^2$

1b optimize $c_j = \frac{1}{|X_j|} \sum_{x \in X_j} x$

fix X_j (not implicit w/ ϕ_S)

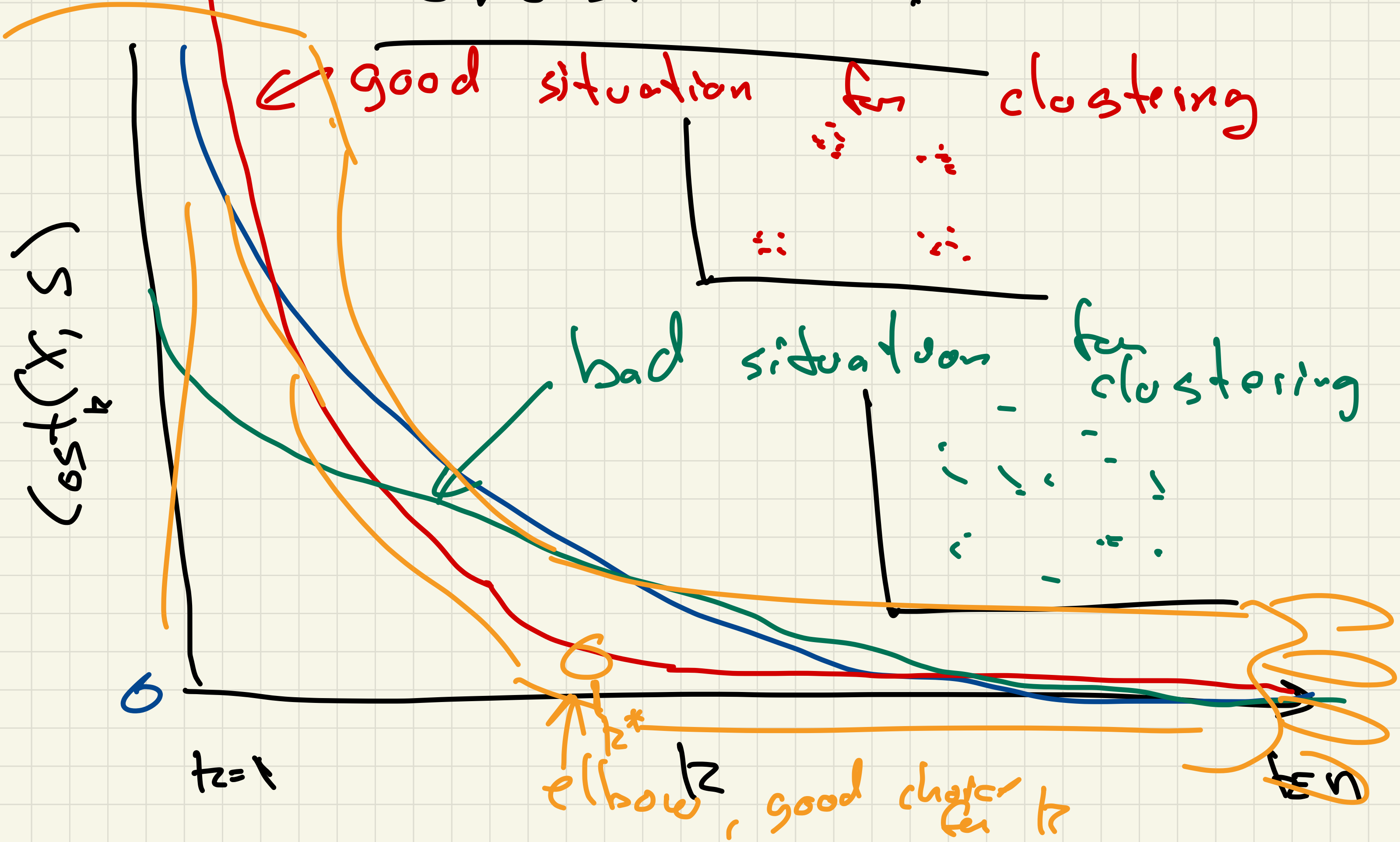


$$h^2 = d_x^2 + d_y^2$$

$$\|x - c_j\|^2 = d_x^2 + d_y^2$$

optimize independently

Choose k ?


good situation for clustering
bad situation for clustering
cost(X, S_k)
0
k=1
k
k=n
k*
elbow, good choice for k