FoDA L23

K-means clustering

↳ Lloyd's Algorithm

Nov 15, 2022

Assignment-based Clustering

Input $X \subset \mathbb{R}^d$ distance $d: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$

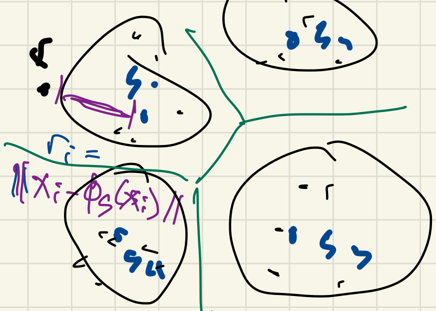
value k $d = \|\cdot - \cdot\|$

\# clusters

k-means

Goal: Set of k sites $S = \{s_1, s_2, \dots, s_k\} \subset \mathbb{R}^d$

$$S^* = \arg\min_{|S|=k} \boxed{\sum_{i=1}^{n} \|x_i - \phi_S(x_i)\|^2} = \mathrm{Cost}(X, S)$$



$\phi_S(x)$ = closest site s_j to x

$$= \arg\min_{s_j \in S} \|s_j - x\|$$

Lloyd's Algorithm

0. **Initialize**: Choose k points $S \subset X$

1. **repeat**

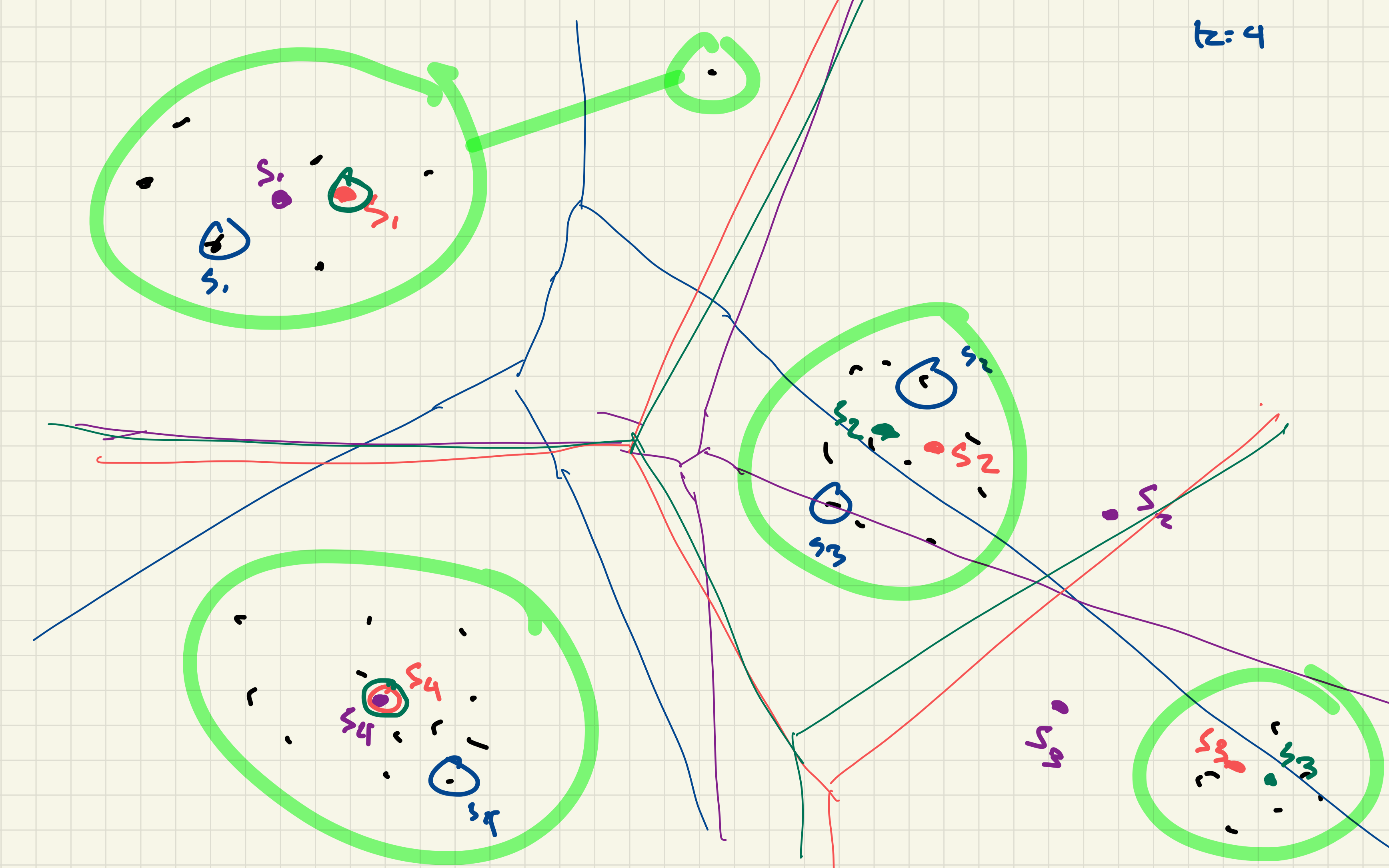
   (a) $\forall\, x_i \in X$: assign x_i to X_j so $\phi_S(x_i) = s_j$

   with subset cluster $X_j \subset X$

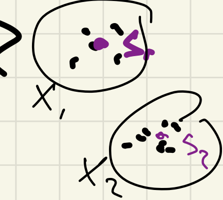
   (b) $\forall\, s_j \in S$: update $s_j = \frac{1}{|X_j|} \sum_{x \in X_j} x$

   until ("converged" or R steps)

   say $R = 10$ or 20


k=4
S1
S1
S1
S2
S2
S2
S2
S3
S3
S3
S3
S4
S4
S4

Does Lloyd's Algo Converge?



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

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

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

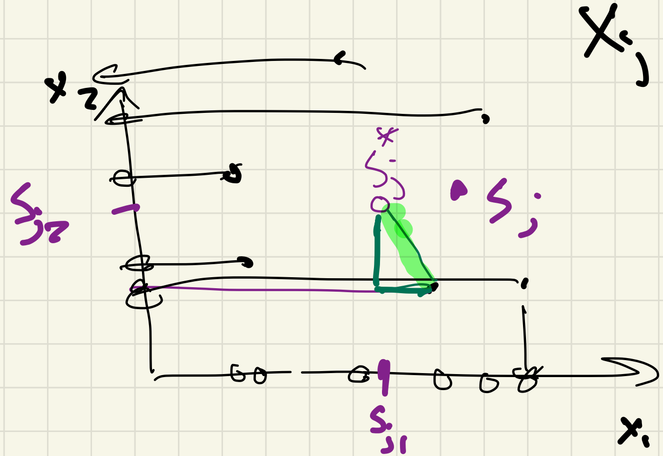
every step ((a) or (b)) decreases Cost(X, S)

Step (a): $\forall x \in X \mid$ assign x to X_j so $\phi_S(x) = s_j$

closest site

step (b): $\forall s_j \in S \mid$ $s_j = \frac{1}{|X_j|} \sum_{x \in X_j} x$

$$s_j^* = \arg\min_s \sum_{x \in X_j} \| x - s \|^2$$



$$\|x_i - s_j\|^2 = (x_{i1} - s_{j1})^2 + (x_{i2} - s_{j2})^2$$

k=4

Lloyd's Algo

not gaurenteed to find optimal solution.

Address non-optimal convergence.

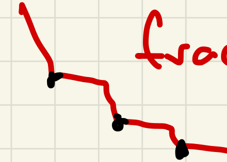
① Try more random initializations "random restart"

② Better Initialization
  (a) Gonzalez (k-centers) (b) k-means++ (randomized)

③ Randomize averaging step
  ↳ simulated annealing

④ Use more clusters $k' \gg k$ ⟶ then merge.
  $k' = k \log n$

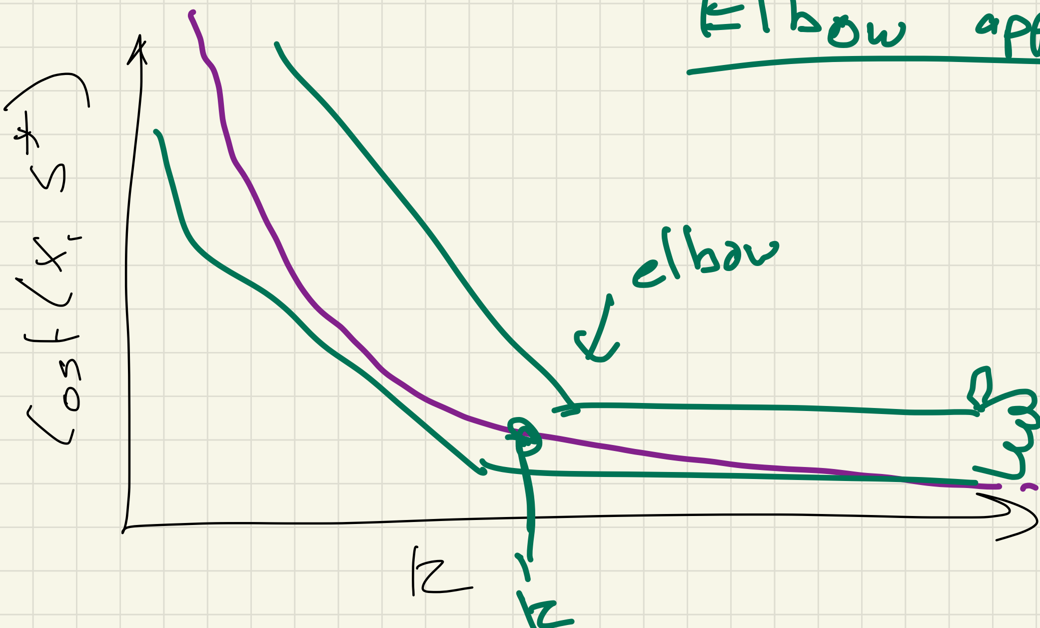
How to choose k ?



frontal

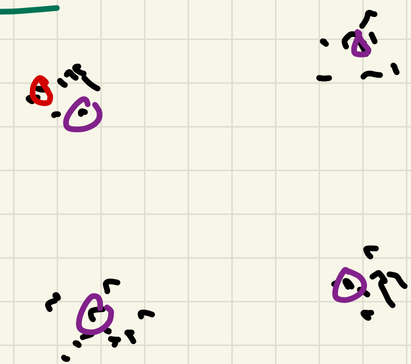
Goal $\quad Cost(X, S) = \sum_{x_i \in X} \| x_i - \phi_S(x_i) \|^2$

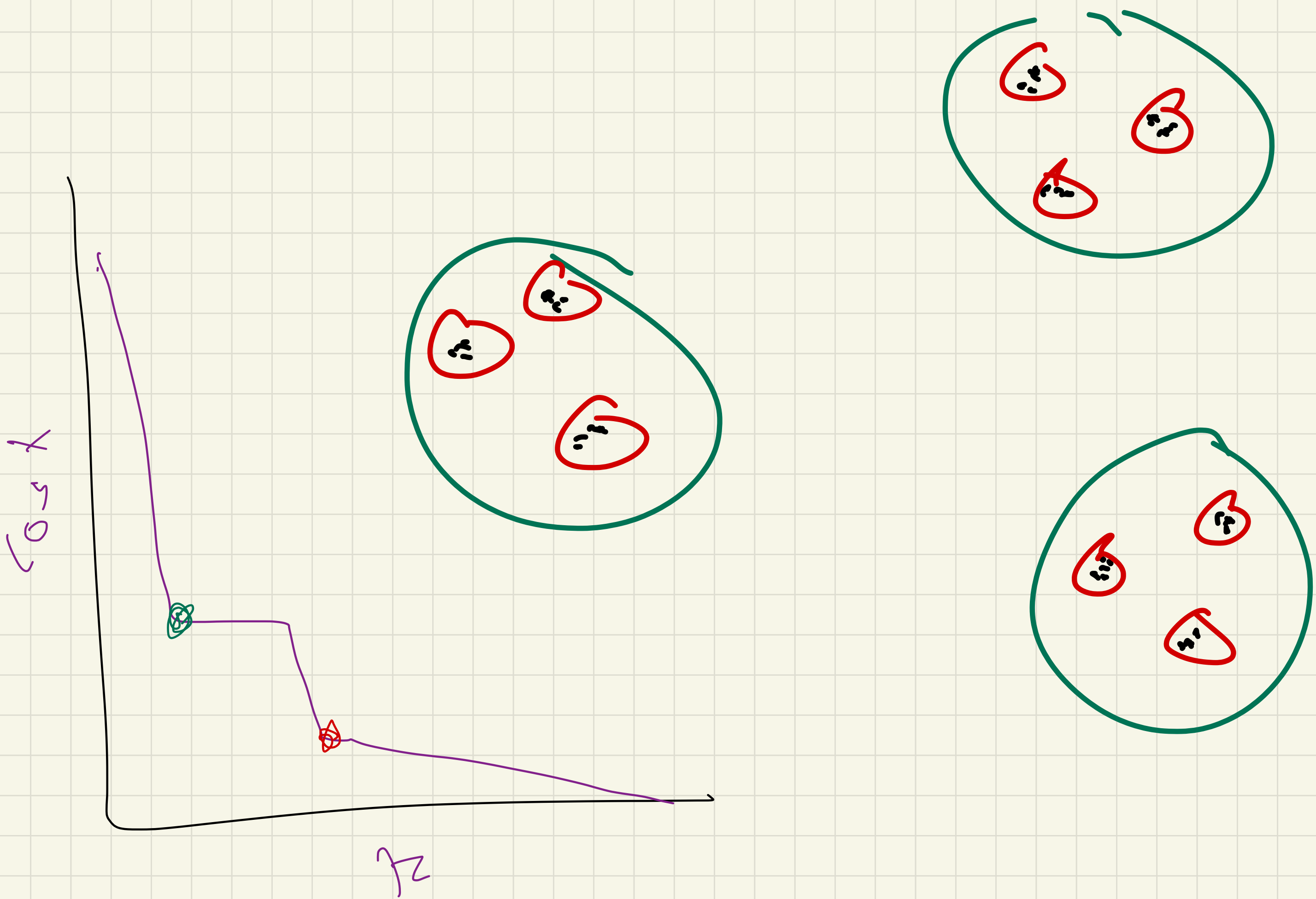
"Elbow approach"

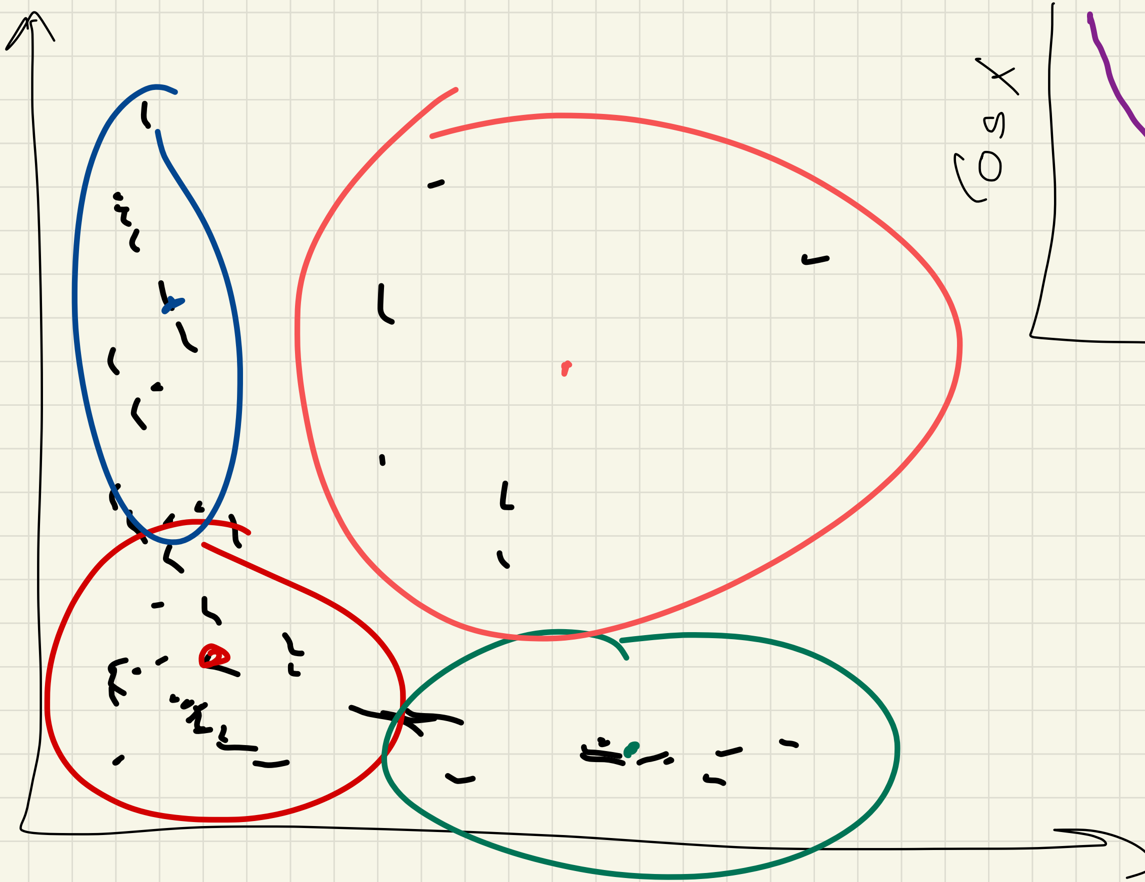


$\to 0 \Rightarrow Cost(X, S^*) = 0$

$k = n$




cost
k




Cost
k
no elbow