FoDA : Clustering

L23 : Lloyd's Algorithm

for k-means

K-Means Clustering Formulation

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

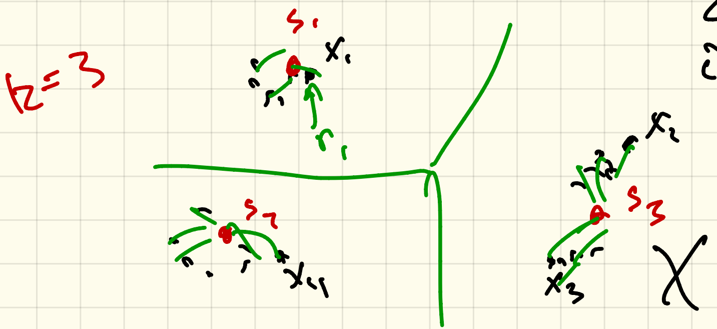
$k \in \mathbb{Z}^+ \in [1, 2, \dots]$

$d : X \times X \to \mathbb{R}_+$ $d(x_1, x_2) = \|x_1 - x_2\|$

Goal Find s.t. $S = \{s_1, s_2, \dots s_k\} \subset \mathbb{R}^d$

minimize $\sum_{i=1}^{n} d\left(x_i, \underline{\phi_S(x_i)}\right)^2$

↳ maps x_i to $\arg\min_{s_j \in S} \|x_i - s_j\|$

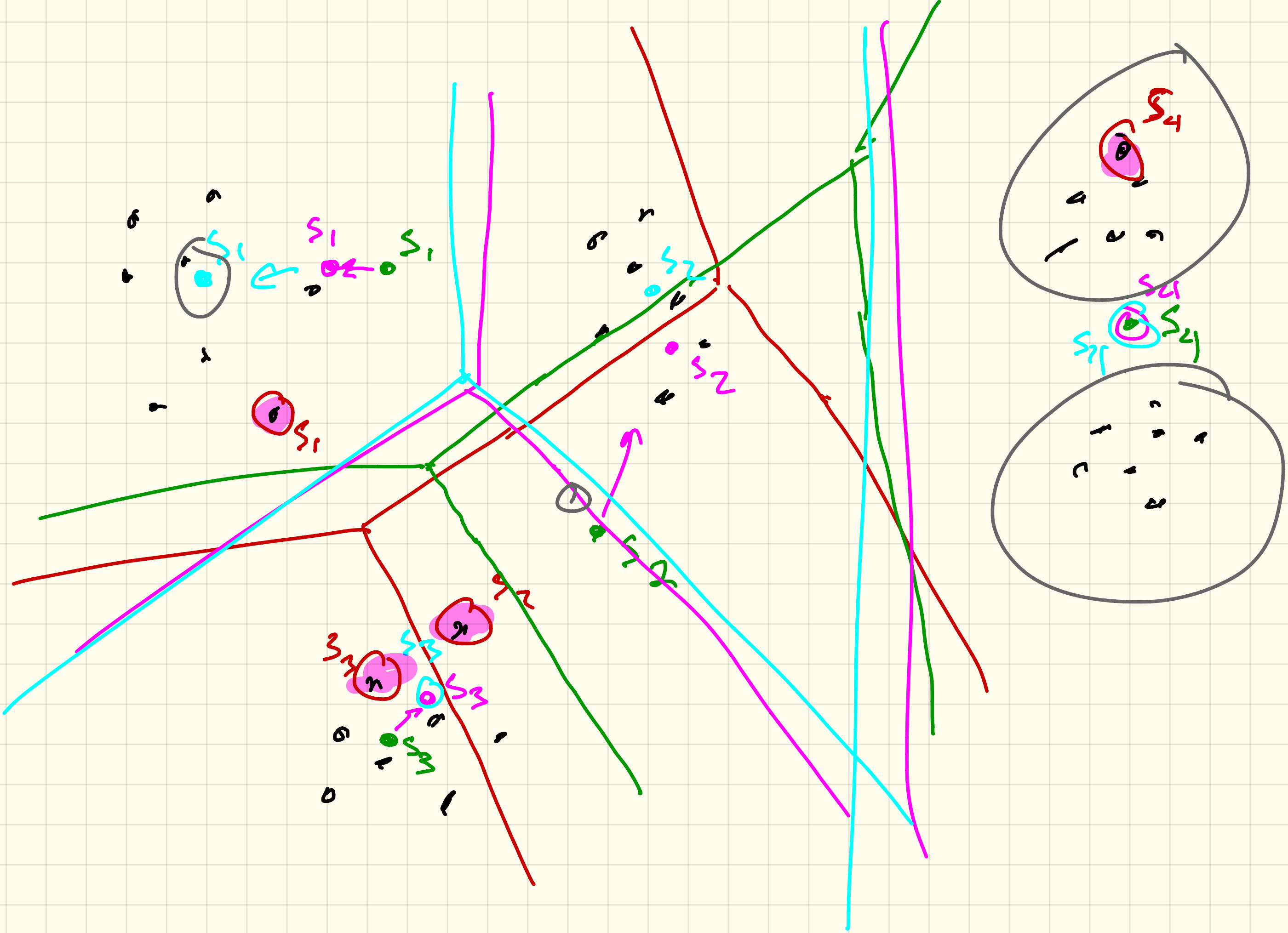


Lloyd's Algorithm → $X = \bigcup_{j=1}^{k} X_j$

$X_j \subset X$

0. Initialize k points $S \subset X$ (arbitrarily?)

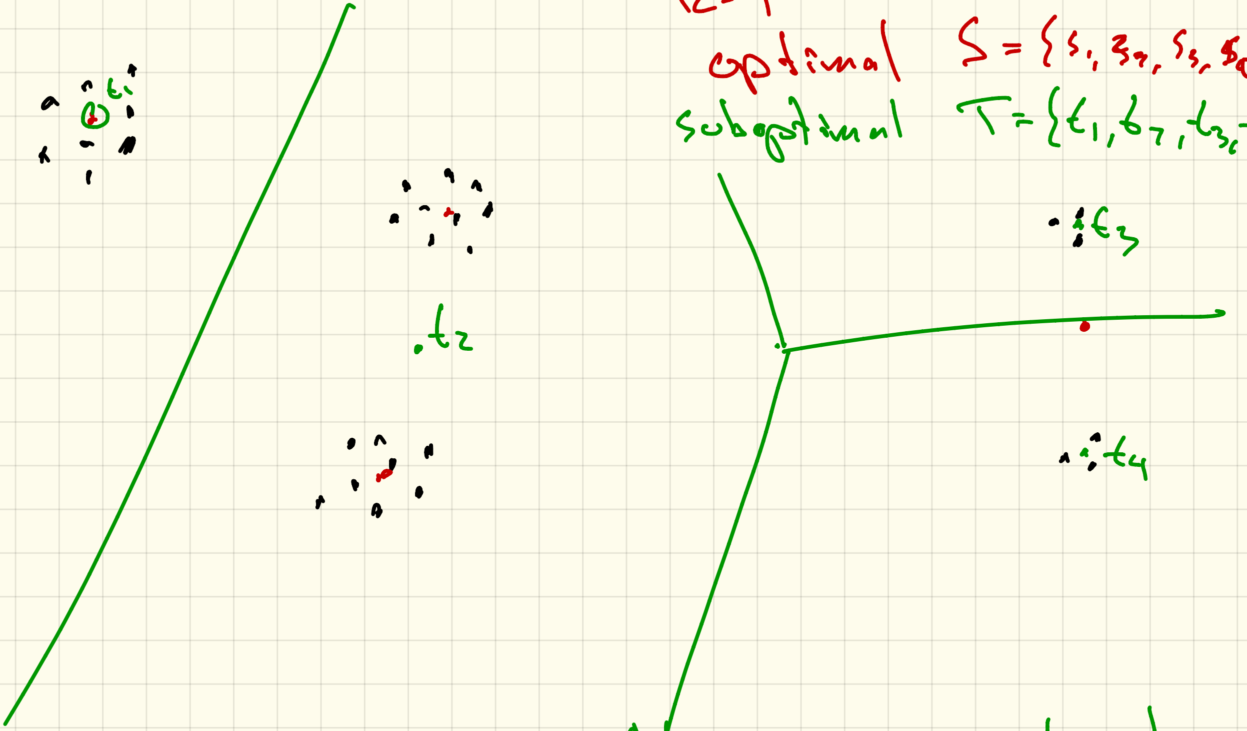
```
1. repeat
   assignment Voronoi dgm  a. for all $x_i \in X$ : assign $x_i$ to $X_j$ so $\phi_S(x_i) = s_j$
   recenter average        b. for all $s_j \in S$ : update $s_j = \frac{1}{|X_j|} \sum_{x \in X_j} x$ = average$(X_j)$
   until (the set S unchanged
          or change is small)
```


S1
S1
S1
S1
S2
S2
S2
S2
S3
S3
S3
S3
S4
S4
S4
S4

$k=4$

optimal $S=\{s_1,s_2,s_3,s_4\}$

suboptimal $T=\{t_1,t_2,t_3,t_4\}$



Lloyd's Algo is <u>stuck</u>

Most of the time Lloyd's works well
(with a little help)

Tricks to help Lloyd's

- Random Restarts

  ×10

  a. Randomly initialize S.

  b. Run Lloyds → compute cost (SSE)

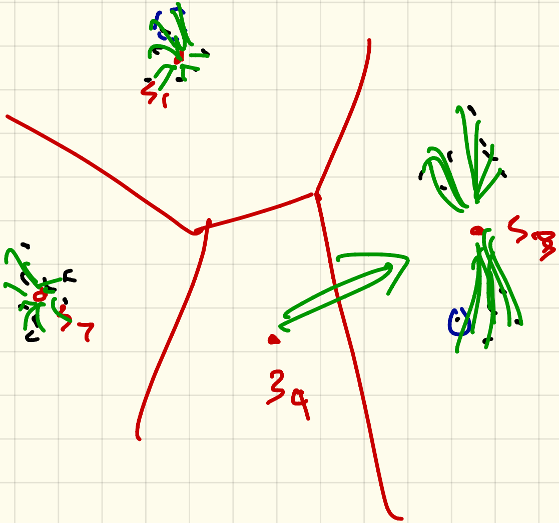
  c. Repeat (a, b) say 10 times

  d. Return final S w/ lowest cost.

- Better Initialize
  - Gonzalez Algo.
  - k-means++

Corner Cases

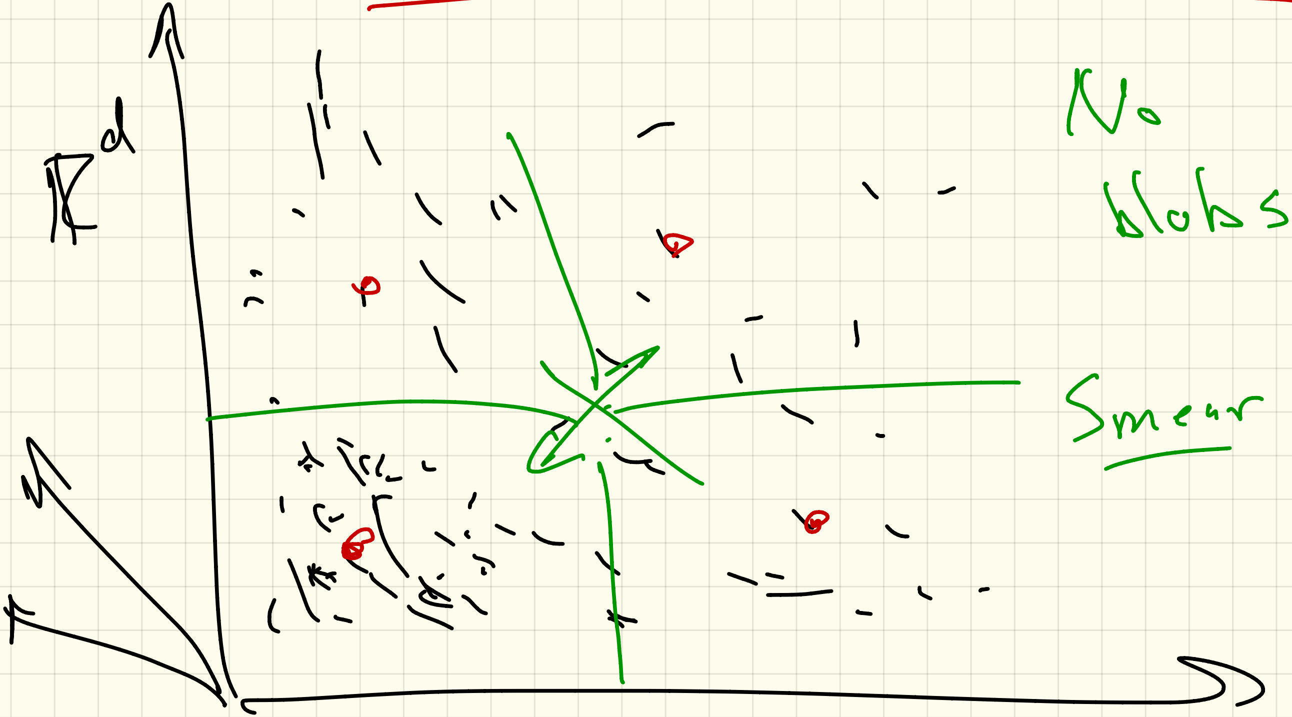
- Might be site w/ no points closest to it



from X_j w/ most variance.

→ Randomly assign s_4 (w/ no points in X_4) to some $x_i \in X$.

Most Data Looks Like

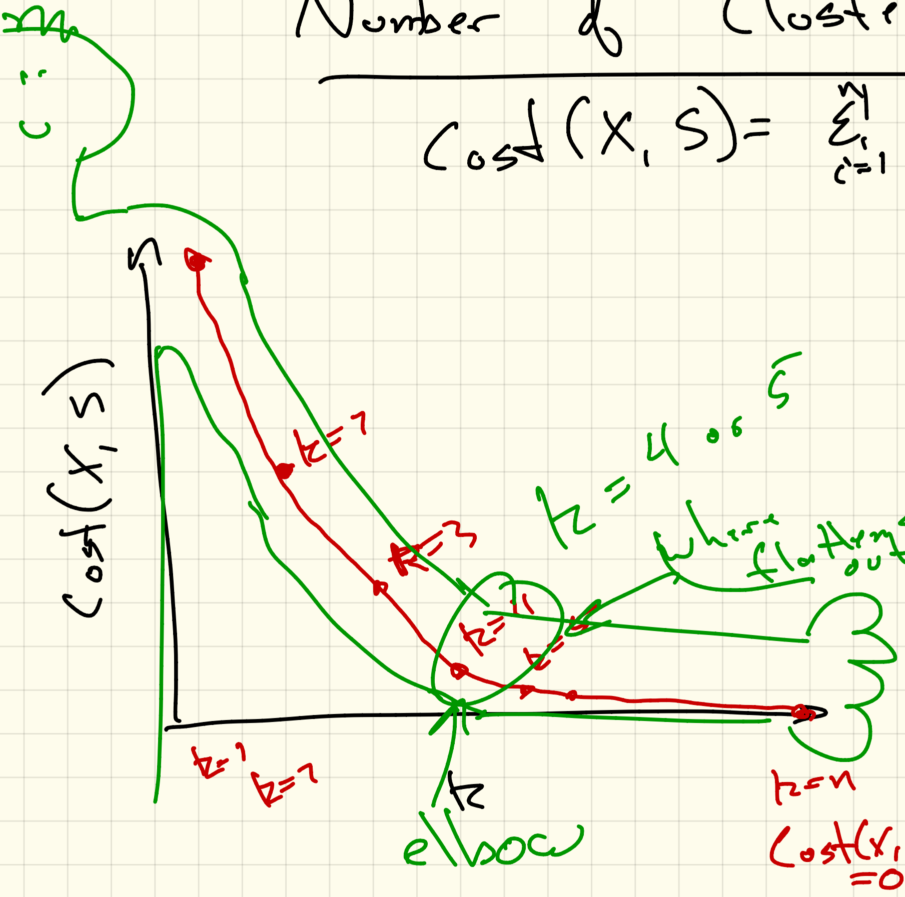


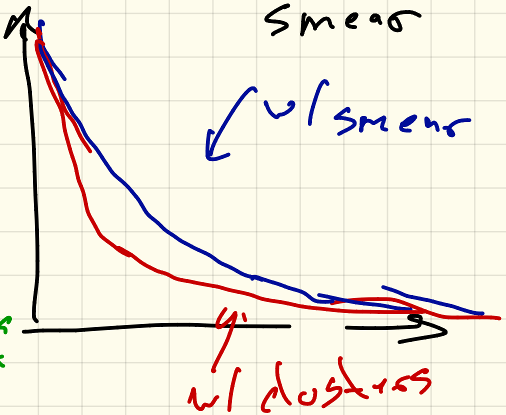
No blobs!

Smear

Number of Clusters. k?

$$\text{Cost}(X, S) = \sum_{i=1}^{n} \| x_i - \phi_S(x_i) \|^2$$





Why Lloyd's Algo converges?

$$\mathrm{cost}(X, S) = \sum_{x \in X} \| x - \phi_S(x) \|^2 \quad \leftarrow (a)$$

$$= \sum_{\underline{s_j \in S}} \left[\sum_{x \in X_j} \| x - s_j \|^2 \right] \quad \leftarrow (b)$$

(a) assignment

(b) centering