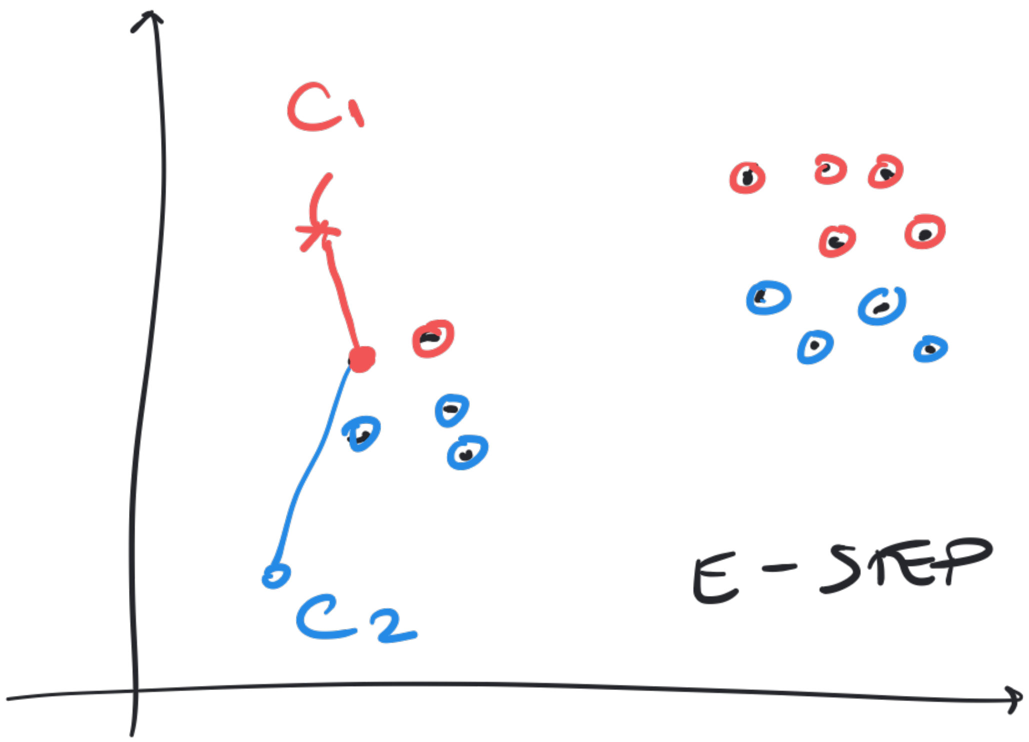
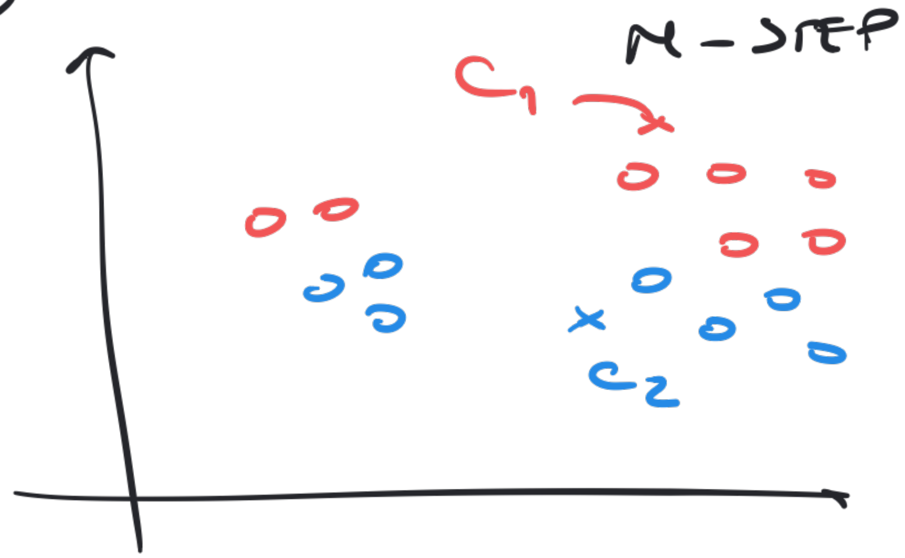
K = 2



Clustering

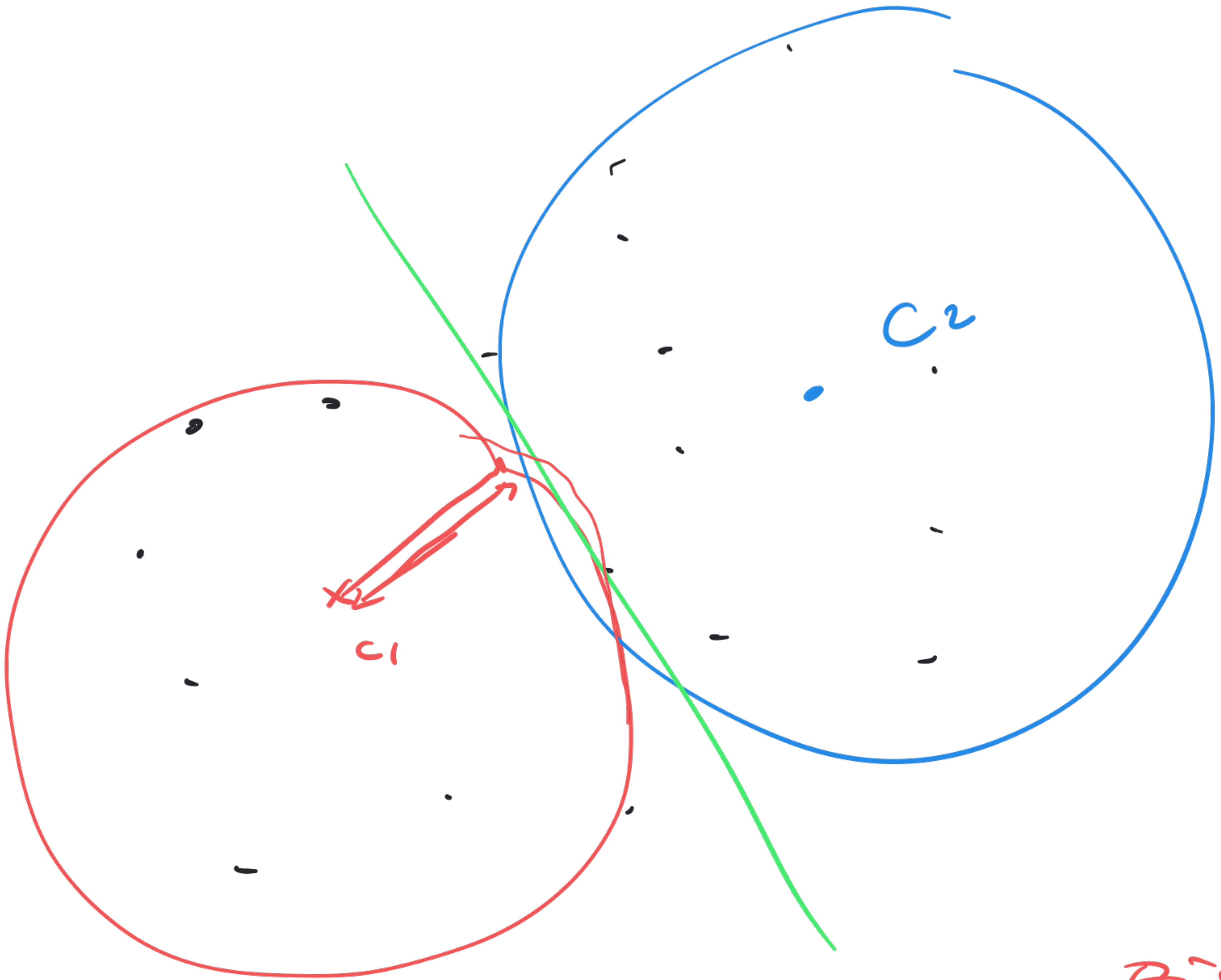
unsupervised learning algorithm



K-Means Clustering

→ K is the # of groups (hyperparameter)

→ centroid-based clustering



Euclidean Distance

when we use as similarity measure, clusters will be spherical/circular.

other distance metrics :

1. city-block distance — measures dist. parallel to canonical axis
2. cosine distance — measures angle between samples.
3. MAHALANOBIS distance — introduces covariance

K-Means uses Alternating optimization which means that its solution will depend on initial conditions for C_K, r_K.

Assignment in K-Means is known as a Hard Assignment

So each point can <u>only</u> belong to a <u>single cluster</u>

<u>Membership Matrix</u> : U , $N \times K$

$$U = \begin{array}{c c} & \begin{array}{cccc} 1 & 2 & \cdots & K \end{array} \\ \begin{array}{c} 1 \\ 2 \\ \vdots \\ N \end{array} & \left[\begin{array}{cccc} 0 & 1 & \cdots & 0 \\ & & & \\ & & & \\ & & & \end{array} \right] \end{array}$$

$$\sum_{j=1}^{K} u_{ij} = 1, \forall i$$

<u>SPARSE MATRIX</u>

$u_{ij} \equiv$ membership of point x_i in cluster with centroid θ_j

$\theta_j \equiv$ centroid for cluster j

$d(x_i, \theta_j) \equiv$ distance between x_i and θ_j

e.g. Euclidean distance : $d(x_i, \theta_j) = \| x_i - \theta_j \|_2$

$$= \sqrt{\sum_{l=1}^{d} \left(x_{i(d)} - \theta_{j(d)}\right)^2}$$

OBJECTIVE FUNCTION:

$$\boxed{J(\theta, U) = \sum_{i=1}^{N} \sum_{j=1}^{K} u_{ij} \cdot d^2(x_i, \theta_j)}$$

Constraints : $\sum_{j=1}^{K} u_{ij} = 1$ and $u_{ij} \in \{0, 1\}$

$$\boxed{\underset{\{\theta, U\}}{\arg\min}\ J(\theta, U)}$$