Q1 'N' data points, 'K' components / clusters.

initialization :- pick 'K' centroids from dataset :
$\mu_1, \mu_2, \ldots, \mu_K$

OBJECTIVE :- $\min \sum \text{dist}(x_i, \mu_K) \cdot \pi_{iK}$

$= \min \sum \|x_i - \mu_K\|^2 \cdot \pi_{iK}$ (for Euclidean dista

Algorithm :-

LOOP (until convergence $\Leftrightarrow$ no change)

E step — Compute membership (given μ) $\pi_{iK} = \begin{cases} 1 & \text{if } x_i \in \text{cluster } K \\ 0 & \text{otherwise} \end{cases}$

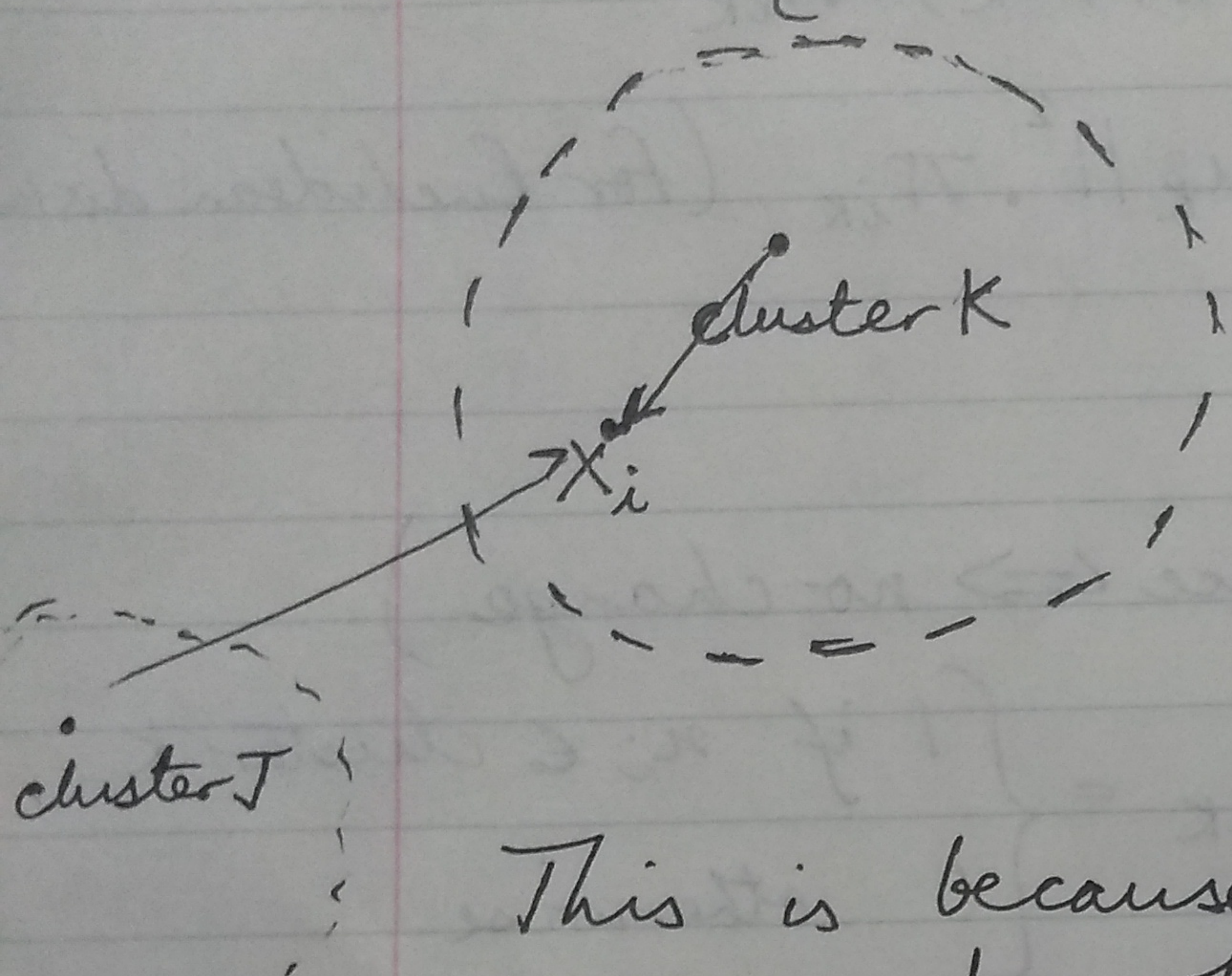
$\pi_{iK} = 1 \Leftrightarrow K = \underset{K}{\text{argmin}} \ \text{dist}(x_i, \mu_K$

M step — Compute centroids (given π) $\mu_K = \text{avg}(x_i \in \text{cluster } K)$

$$= \frac{\sum_i x_i \cdot \pi_{iK}}{\sum_i \pi_{iK}}$$

A) To obtain the minimum global objective, in E step, the only way to achieve it is if the membership π_{ik} satisfies the equation:

$$\pi_{ik} = \begin{cases} 1 \text{ if } x_i \in \text{cluster } K \\ 0 \text{ otherwise.} \end{cases}$$



This is because, if X_i belongs to or is a member of any other cluster, ~~distance~~ ~~distance of~~ say cluster J, then, distance (X_i, μ_j) will not be minimum. and hence, objective function will not be minimum.

B) Differentiating $\sum \|x_i - \mu_K\|^2 . \pi_{iK}$ w.r.t μ_K,

$$\frac{d}{d\mu_K}\left(\frac{\sum X_i . \pi_{iK}}{\sum \pi_{iK}}\right) = 0$$, so this will be a

minima, and hence $\mu_K = \text{avg}(x_i \in \text{cluster } K)$ will achieve minimum objective.

c) There are at most K^N ways to partition 'N' datapoints into K clusters. This is a large but a finite number. For each iteration, we produce a new clustering based only on the previous old clustering.

1) If old clustering is same as the new, next clustering will again be same.

2) If new clustering is different from old clustering then newer one has lower cost.

Since the algorithm iterates a function whose domain is a finite set, the iteration must eventually enter a cycle. The cycle cannot have length > 1, because otherwise by (2) you would have some clustering which has lower cost than itself, which is a contradiction & hence not possible. Hence cycle must have exactly length = 1. Hence K-means converges is a finite no. of iterations.