2. Convergence of K-Means

Given a set of n point $p_i \in \mathbb{R}^d, i \in \{1, 2, \ldots, n\}$ and the number of clusters k, the K-means clustering algorithm aim to find the centers of k clusters $c_j, j \in \{1, 2, \ldots, k\}$ by minimizing the average distance from n points to their assigned closest cluster centers. The loss function to be minimized can be formulated as:

$$L(c) = \sum_{i=1}^{n} \min_{j \in \{1,\ldots,k\}} ||p_i - c_j||_2^2 \tag{1}$$

To approximate the solution, the new assignment variables $z_i \in argmin_{j \in \{1,\ldots,k\}} ||p_i - c_j||_2^2$ for each data point p_i is introduced. The K-means clustering algorithm iterates between updating the variables z_i (*assignment step*) and updating the centers $c_j = \frac{1}{|\{i:z_i=j\}|} \sum_{i:z_i=j} p_i$ (*refitting step*). The algorithm stops when no change occurs during the *assignment step*.
Please prove that K-means is guaranteed to converge (to a local optimum). Note: You need to prove that the loss function is guaranteed to decrease monotonically in each iteration until convergence. Prove this separately for the *assignment step* and the *refitting step*, provide your solution in a pdf file.
(3 Points)

Solution:

To prove convergence of the K-means algorithm, we show that the loss function is guaranteed to decrease monotonically in each iteration until convergence for the assignment step and for the refitting step. Since the loss function is non-negative, the algorithm will eventually converge when the loss function reaches its (local) minimum.

Let $z = (z_1, \ldots, z_n)$ denote the cluster assignments for the n points.

(a) Assignment step

We can write down the original loss function $L(c)$ as follows:

$$L(c, z) = \sum_{i=1}^{n} ||p_i - c_{z_i}||_2^2 \tag{2}$$

Let us consider a data point p_i, and let z_i be the assignment from the previous iteration and z_i^* be the new assignment obtained as:

$$z_i^* \in \arg \min_{j \in \{1,\ldots,k\}} \|p_i - c_j\|_2^2$$

Let z^* denote the new cluster assignments for all the n points. The change in loss function after this assignment step is then given by:

$$L(c, z^*) - L(c, z) = \sum_{i=1}^{n} \left(\left\|p_i - c_{z_i^*}\right\|_2^2 - \|p_i - c_{z_i}\|_2^2 \right) \leq 0$$

The inequality holds by the rule z_i^* is determined, i.e. to assign p_i to the nearest cluster.

(b) Refitting step

We can write down the original loss function $L(c)$ as follows:

$$L(c, z) = \sum_{j=1}^{k} \left(\sum_{i:z_i=j} \|p_i - c_j\|_2^2 \right)$$

Let us consider the j^{th} cluster, and let c_j be the cluster center from the previous iteration and c_j^* be the new center obtained as:

$$c_j^* = \frac{1}{|\{i : z_i = j\}|} \sum_{i:z_i=j} p_i$$

Let c^* denote the new cluster centers for all the k clusters. The change in loss function after this refitting step is then given by:

$$L(c^*, z) - L(c, z) = \sum_{j=1}^{k} \left(\left(\sum_{i:z_i=j} \left\|p_i - c_j^*\right\|_2^2 \right) - \left(\sum_{i:z_i=j} \|p_i - c_j\|_2^2 \right) \right) \leq 0$$

This inequality holds because the update rule of c_j^* essentially minimizes this quantity.